Nonlinear breathing modes at a defect site in DNA

Ciprian-Ionuț Duduială,[1] Jonathan A. D. Wattis,[1] Ian L. Dryden,[1] and Charles A. Laughton[2]
[1] *School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom*
[2] *School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, United Kingdom*

(Received 3 March 2009; revised manuscript received 9 September 2009; published 10 December 2009)

Molecular-dynamics simulations of a normal DNA duplex show that breathing events typically occur on the microsecond time scale. This paper analyzes a 12 base pairs DNA duplex containing the "rogue" base difluorotoluene (F) in place of a thymine base (T), for which the breathing events occur on the nanosecond time scale. Starting from a nonlinear Klein-Gordon lattice model and adding noise and damping, we obtain a mesoscopic model of the DNA duplex close to that observed in experiments and all-atom molecular dynamics simulations. The mesoscopic model is calibrated to data from the all-atom molecular dynamics package AMBER for a variety of twist angles of the DNA duplex. Defects are considered in the interchain interactions as well as in the along-chain interactions. This paper also discusses the role of the fluctuation-dissipation relations in the derivation of reduced (mesoscopic) models, the differences between the potential of mean force and the potential energies used in Klein-Gordon lattices, and how breathing can be viewed as competition between the along-chain elastic energy and the interchain binding energy.

DOI: 10.1103/PhysRevE.80.061906 PACS number(s): 87.15.H−, 05.45.−a, 87.14.gk, 87.10.Tf

I. INTRODUCTION

This paper studies the nucleation of open bubbles in *deoxyribonucleic acid* (DNA). We show that breathing can start at a defect site of the DNA sequence; this is motivated by the results of Cubero *et al.* [1]. We show that the defect not only weakens the hydrogen bonds between complementary base pairs but also changes the along-chain stacking interactions and so alters the breathing behavior of the DNA sequence. Since other macromolecules which DNA interacts with may alter the twist angle of the double helix, we investigate how twist influences the frequency and duration of breathing events by proposing a model based on a system of stochastic differential equations, with parameters fitted to data obtained using the well-established molecular dynamics (MD) package AMBER. We show how the along-chain and interchain interaction parameters vary with local twist. This twist is imposed through external constraints at the extremities of the DNA sequence in a similar fashion to the twist perturbations imposed by DNA-binding proteins.

DNA is a nucleic acid whose main role is the long-term storage of genetic information needed for the development and functioning of living organisms. From a structural point of view, DNA is a long polymer composed of simple units called nucleotides, which are held together by a backbone of sugars and phosphate groups. The nucleotides composing a DNA sequence differ in their bases, which encode the genetic information copied by cells from DNA into RNA in order to use. These bases are of four types from two different categories: the purines adenine (A) and guanine (G)—having two organic cycles—and the pyrimidines cytosine (C) and thymine (T)—with only one organic cycle.

Watson and Crick [2] first introduced, in 1953, the molecular structure of a DNA sequence. A DNA duplex is composed of two chains of bases. A base from one chain has a corresponding base on the other chain which together form a so-called *base pair*. Adenine (A) forms a base pair with thymine (T), while guanine (G) pairs with cytosine (C). The bases are linked by covalent bonds along the chains, while the bases of each pair are linked together as follows: A-T pairs by two hydrogen bonds and C-G pairs by three hydrogen bonds [3]. In addition, the double-stranded DNA is twisted around its central axis. The twist is typically 36° per base pair. Using this information, computer simulations of the DNA structure can be carried out at different levels of resolution [4].

One of the techniques to investigate DNA processes is MD simulations using computer programs, such as AMBER. The biggest inconvenience with such an approach is the time spent simulating. The molecule cannot be analyzed alone as the solvent surrounding the DNA, which in our case is water, needs to be taken into account. For this reason, during MD simulations a lot of time is spent analyzing the solvent containing many times more atoms than the DNA sequence under study, resulting in the overall time needed for just one simulation to be of weeks or months even when several processors work in parallel. This is why a simplified dynamic model of DNA is needed.

Recently, mathematical models of processes that take place in a DNA sequence have been developed. These models—using linear [5], nonlinear [6], or geometrical approaches [7]—can be used to predict the behavior of DNA or to analyze measurable quantities, for example, the system's energy. Many of these models study DNA denaturation and unzipping (see, for example, [8]), but they can also be used to analyze breathing modes. A breathing event represents the opening of one or more base pairs. In other words, it means the temporary breaking of the hydrogen bonds between complementary bases. Such events can be examined at both macroscale or microscale. Experimental work of Altan-Bonnet *et al.* [9] analyses both the initiation of base-pair opening and the growth of open bubbles in double-stranded DNA. They observe kinetics on a wide range of time scales. Once a bubble has formed, they model the process stochastically using constant rates of growth and shrinkage. A similar model of bubble growth is presented by Ambjörnsson *et al.* [10], who considered the sequence dependent aspects of the problem.

Simple linear and nonlinear mechanical models allow relevant modes to be analyzed. Salerno [11] suggested that sine-Gordon kinks are set in motion in certain regions of a DNA sequence that includes promoters. The proposed model analyzes nonlinear wave dynamics of the $T7A_1$ DNA promoter and is based on the following equations of motion:

$$I\frac{d^2\psi_i}{dt^2} = K(\psi_{i+1} - 2\psi_i + \psi_{i-1}) - \frac{\beta}{2}\lambda_i \sin(\psi_i - \theta_i),$$

$$I\frac{d^2\theta_i}{dt^2} = K(\theta_{i+1} - 2\theta_i + \theta_{i-1}) - \frac{\beta}{2}\lambda_i \sin(\theta_i - \psi_i),$$

where θ_i and ψ_i represent the deflection angles that two complementary bases form with the imaginary line connecting them, while K is the backbone spring constant, I is the moment of inertia of a base, β is a parameter, describing the strength of the base-pair interaction, and λ_i represents the number of hydrogen bonds involved in pairing the bases (λ_i=2 or 3, depending on whether the base pair is A-T or C-G, respectively).

The DNA model proposed by Muto *et al.* [12] considers the two polynucleotide strands to be springs, while the bases of a pair are linked together by hydrogen bonds, which are described by a Lennard-Jones potential. They obtain, from the equations of motion, the expressions for the transverse and longitudinal displacements of each base. Even when studies of DNA sequences are based on multidimensional models, such as [12], most models are reduced to a one-dimensional system by taking into account only the transverse displacements and describe how the distances between paired bases vary in time instead of computing the actual position of each base. Van Zandt, for example, analyzed only the transverse displacements [13], taking into account both the elastic restoring force between neighbors on the same strand and an intrastrand force between complementary bases. Such models emphasize the links with breather modes and other solitons.

Peyrard and Bishop [6] proposed one of the first nonlinear models, which neglects the inhomogeneities due to the base sequence and the asymmetry of the two strands. This model, analyzed in [14,15], ignores the longitudinal displacements, while the neighboring nucleotides of the same strand are connected by a harmonic potential to keep the model as simple as possible. Considering a common mass m for all bases and the same coupling constant k along each strand, they define the system's Hamiltonian as

$$H = \sum_n \frac{1}{2}m\left[\left(\frac{du_n}{dt}\right)^2 + \left(\frac{dv_n}{dt}\right)^2\right] + \frac{1}{2}k[(u_n - u_{n-1})^2 + (v_n - v_{n-1})^2] + V(u_n - v_n),$$

where u_n and v_n represent the nth bases' displacements from equilibrium. The nonlinearity is introduced via the Morse potential $V(u_n - v_n) = D(e^{-a(u_n - v_n)} - 1)^2$, with D and a being the depth and the inverse width of the Morse potential. This potential describes the bonds connecting the opposite parts of a base pair, which are stretched when the double helix opens locally. Using a similar model, Peyrard and Farago [16] proved that, at low temperatures, localization is due to individual discrete breathers, while, at high temperatures, large regions are involved.

Barbi *et al.* developed a new model with two degrees of freedom per base pair [17]. They studied analytically small amplitude dynamics of the model, in which the bases are allowed to move in the base plane described by a radial variable specific to the motion along the hydrogen bonds and an angular variable indicating the base-pair twisting degree. In a recent paper [18], Gaeta and Venier identified the conditions for which solitary traveling waves exist in the model of Barbi *et al.*

In [19,20] Ambjörnsson and Metzler used the dynamic approaches, based on a (2+1)-dimensional master equation and a Fokker-Planck equation, respectively, to study the size fluctuations of bubbles in a DNA molecule in the presence of single-stranded DNA-binding proteins. Hanke and Metzler [21] studied the bubble dynamics of double-stranded DNA using a Fokker-Plank equation based on the bubble's free energy function, which allows then to include microscopic interactions in a straightforward fashion. Another scheme, based on a stochastic approach and describing temporal fluctuations of local denaturation zones in double-stranded DNA, is proposed by Banik *et al.* [22]. In fact, stochastic approaches may represent a mesoscopic model for long time scale simulations of long chains, which are inaccessible to all-atom molecular dynamics studies. During breathing events, random oscillations of the base-pair displacements are observed. It is possible to model these events using stochastic processes, but it is not clear how to calibrate such models to data from experiments and MD simulations of DNA. Another problem is that random terms can increase considerably the temperature and the total energy of the system. Lennholm and Hornquist [23] presented the Nosé-Hoover thermostat as the simplest version of such a model. This approach introduces an extra degree of freedom into the system, which has the role of maintaining the temperature at a certain value. Quintero *et al.* [24] also used a stochastic approach introducing a damping term into the system, which conserves energy. Such an approach relates the temperature to the noise terms that simulate the random events in the system.

Taking all these aspects into account, mesoscopic DNA models are still a challenge for nonlinear science, as discussed by Peyrard *et al.* in [25]. The main challenge is the choice of the potentials describing the interactions from the system. For interstrand interactions Zhang *et al.* [26] analyzed the Toda lattice potential and the Morse potential. Using a transformation of variables and the Morse potential they showed that a solitary wave excitation with an estimated width of only one or two base pairs can be obtained. Peyrard *et al.* [25] suggested that the simple Morse potential is not enough to describe all the DNA effects and proposed a more elaborate function containing a barrier for reclosing base pairs. The stacking interactions (between bases situated on the same chain) are also important in such systems. Most of the papers consider harmonic coupling along strands, but in [25,27,28] it is suggested that nonlinear stacking leads to a self-amplification process. The improved stacking potential has the role of weakening the along-chain bonds during a breathing event. This lengthens the breathing events since it causes a weaker closing force. However, a choice of along-chain and interchain potentials that allows the profiles of wave excitations to be found does not guarantee that the DNA behavior is accurately represented by these mathematical models unless simulations can show that the results are close to experimental data or all-atom molecular dynamics simulations.

Base-pair breathing in DNA typically occurs on the microsecond time scale [3], which is beyond the scope of all-atom molecular dynamics simulations. Replacing a thymine with a difluorotoluene base force breathing occurs on the nanosecond time scale, which allows us to study such events using both MD simulations and other methods. Guckian *et al.* [29] discussed the properties of a 12-mer duplex having a thymine base (T) replaced with the "rogue" base difluorotoluene (F). They concluded that the geometry of the Watson-Crick model is not affected by this change, but that it leads to the formation of weak hydrogen bonds between the A-F base pair. More precisely, only one hydrogen bond links the adenine (A) to the nonpolar molecule (F), weakening the interchain interaction at this defect point in DNA. A similar result was derived by Wattis *et al.* [5] and used in the model studies in [30]. Several studies consider DNA sequences with such a defect to be a probe for the DNA replication mechanism—see [31], for example, in which it is suggested that conventional hydrogen bonds are not crucial for high efficiency and fidelity in DNA synthesis. Moreover, in DNA strands which incorporate a defective base, DNA breathing has been observed to occur on the nanosecond time scale, as presented in a recent study made by Cubero *et al.* [1].

In this paper, we focus only on stationary breathers appearing at the defect site of the considered lattice. Our molecular dynamics simulations, obtained using AMBER [32], revealed that the frequency, amplitude, and duration of breathing events vary with helical twist but in a complex way that at present we do not fully understand. We therefore sought to see if a simpler model, with fewer variables, could reproduce this twist-dependent behavior, in which undertwisted DNA (30°–35° per base pair) displays frequent short-duration breathing events, while in overtwisted DNA (37°–40° per base-pair) longer-duration breathing can be observed. We therefore propose a stochastic mesoscopic model for this behavior, fitted to MD data, and we compare the results to all-atom AMBER simulation.

After briefly reviewing the AMBER package, in Sec. II, we present the results of our MD simulations which show the twist-breathing dependence and the need of a reduce model to explain it. Next, in Sec. III we consider each base as a single particle and, using a change of variables, reduce the model to one dimension. We also introduce noise and damping terms into our system and show the need of an alternative fluctuation-dissipation relation for reduced models. In Sec. IV we use the maximum likelihood estimation (MLE) method to fit the unknown parameters to data obtained from AMBER simulations and we determine an expression for the "potential of mean force" (PMF) of our stochastic differential equation (SDE) system. Next, in Sec. V we use the implicit midpoint method [33] to simulate the breathing process in a 12 base-pair DNA sequence using our SDE model. This section also includes a discussion of the results of parameter fitting, noting how the harmonic stacking potential, the interchain potential (determined from the free energy of the breathing pair), and the damping amplitude all influence the time spent breathing for different twist angles. Finally, Sec. VI includes the conclusions we draw.

II. MICROSCOPIC SIMULATIONS

In some recent MD studies, we have examined how the breathing behavior of F-containing DNA sequences is affected by the twist of the helix. Our alternative to experimental investigation of a double-stranded DNA sequence is the MD-simulation package AMBER, which was initially developed by a group lead by Peter Kollman. In his memory, Case *et al.* continued to develop this computer program [32]. Version 9 has been used for our simulations.

Our AMBER simulation considers a 12-mer DNA sequence solvated in a water box. As already presented, replacing the natural base thymine with the non-natural base difluorotoluene can be used as a probe for DNA breathing. We introduced this defect at one of the middle sites of our sequence and we simulated the system for eight different DNA twist angles in the range of 30°–40°. Since it is not possible to produce "relaxed" models of the same DNA sequence with differing twists, the desired twist is imposed by external constraints at the extremities, approximately mimicking how DNA-binding proteins can also perturb the twist.

AMBER considers one degree of freedom for each atom and computes their coordinates and velocities at each time step. The system is simulated using one of AMBER's tools called SANDER, which requires three input files: a topology file, containing information about each atom and several flags, representing bond values or solvent box dimensions, for example, a coordinates file specifying the initial position of each atom, and a configuration file stating the control variable values needed to determine the type of the simulations to be processed. About 2 weeks and four parallel processors are needed to simulate 20 ns for a system containing 16 682 atoms—out of which only 763 are the base-paired atoms and the rest represent the water box; we use a 2 fs time step.

AMBER data

The results obtained suggest that breathing depends on the helical twist. Figure 1 presents the way in which the distance between the bases of the breathing pair varies over time for a 30° undertwisted strands of DNA. A common way of visualizing and interpreting these data is to use a histogram to classify the amount of time spent at each displacement from equilibrium. Such a representation is shown in Fig. 10 for bins of width 0.5 Å. We observe three different values around which this distance oscillates:

(1) 0 Å—represents the equilibrium (closed or non-breathing) state;

(2) 1.9 Å—represents the first breathing state; and

(3) 3.8 Å—represents the second breathing state.

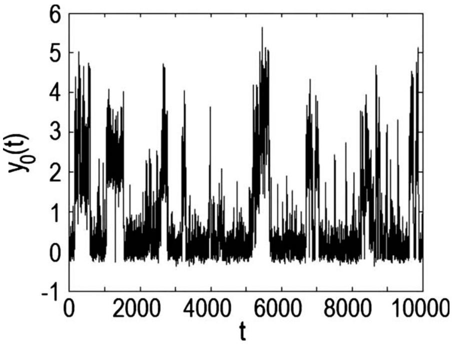

FIG. 1. Graph of the displacement in Å between bases of the breathing pair, plotted against time (ps), obtained from an AMBER simulation of 10 ns, for a 30° undertwisted DNA sequence.

As far as we are aware, it has not yet been determined whether the two breathing states have similar or different causes. One possible explanation is that in the first breathing state only one base of a pair is breathing, while is the second state both bases are breathing.

Analyzing the DNA sequence for the 33° twist angle, we observe in Fig. 2 the same three states explored by the breathing pair. While the time spent breathing is almost the same as in the previous example, the behavior of the DNA sequence is different: the breathing events are longer and less frequent in Fig. 2 than in Fig. 1. Such differences cannot be detected by the use of histograms for such data, namely, in the comparison of Figs. 10 and 13. It is this reason that in this paper we advocate the use of analytical techniques which are more refined than histograms and the associated potential of mean force.

Figure 3 shows that in an AMBER simulation of the typical twist case 36°, the second breathing state is not reached as often as in the undertwisted case and most of the time spent breathing is in the first state. The corresponding histogram for this twist angle is shown in Fig. 15. This shows that the first breathing state is attained more often by the 36° twist than 33° or 30° (compare with Figs. 10 and 13).

Finally, for a 38° overtwisted DNA sequence, the time spent breathing represents more than 65% of the total time of a simulation, as shown in Fig. 4. An important proportion of this time is spent in the second breathing state, in contrast with the previous two cases. This contrast is reflected further when the histogrammed data shown in Fig. 18 are compared with the other histograms (Figs. 10, 13, and 15).

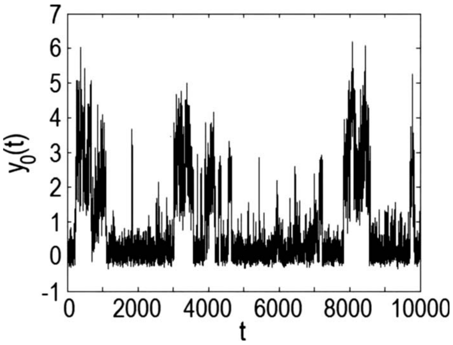

FIG. 2. Graph of the displacement in Å between bases of the breathing pair, plotted against time (ps), obtained from an AMBER simulation of 10 ns, for a 33° undertwisted DNA sequence.

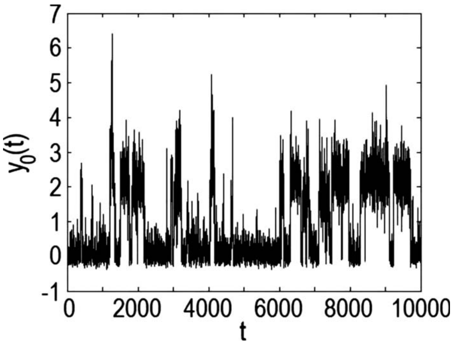

FIG. 3. Graph of the displacement in Å between bases of the breathing pair, plotted against time (ps), obtained from an AMBER simulation of 10 ns, for a 36° twisted DNA sequence.

Analyzing these four simulations, we observe that for an undertwisted DNA sequence the breathing events are short and frequent, while for an overtwisted sequence breathing lasts much longer and is less frequent. The normal twist of 36° represents an average in breathing length and frequency between the undertwisted and overtwisted cases.

III. MESOSCOPIC MODEL

Although AMBER data offer some information about how the twist angle influences breathing, we cannot explain the exact causes, having no information about the along-chain and interchain interactions. For this reason, we construct a stochastic mesoscopic model, which is able to produce simulations close to our MD data. This SDE system has less parameters than AMBER system and allows us to analyze how these parameters vary with twist angle. However, the SDE model that we propose has several parameters and is influenced even by small changes in values of these variables. Hence, we need to fit our parameters to data obtained using AMBER in order to obtain SDE simulations close to the microscopic ones.

A. Derivation

In our model, each base of the DNA sequence is considered to be a separate point mass linked to three other bases: one in each direction along the same chain and one on the

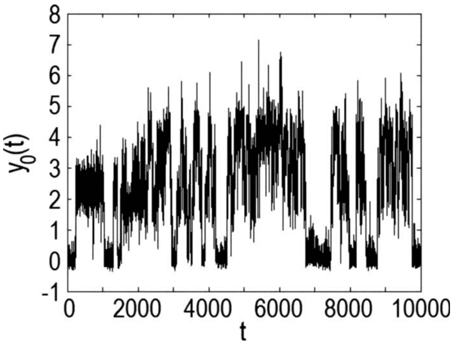

FIG. 4. Graph of the displacement in Å between bases of the breathing pair, plotted against time (ps), obtained from an AMBER simulation of 10 ns, for a 38° overtwisted DNA sequence.

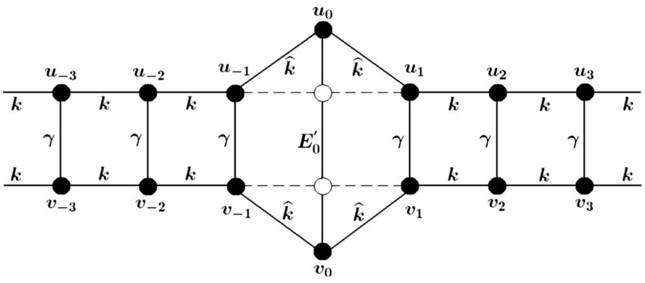

FIG. 5. Illustration of the DNA model.

complementary chain, as shown in Fig. 5. The interchain bonds are modeled by nonlinear force-displacement relationships, while the intrachain bonds are modeled by linear springs with constant k, as in [30]. We construct a model with $4N$ bases—this means $2N$ base pairs—which can be viewed as a lattice of order N. The system is a recursive relation in n, with base-pair $-N$ being identical to base-pair $+N$.

The energy associated with a breathing event is expressed by the Hamiltonian [30]

$$H = \sum_n \frac{1}{2}m_n\left(\frac{du_n}{dt}\right)^2 + \frac{1}{2}m_n\left(\frac{dv_n}{dt}\right)^2 + \frac{1}{2}k^{(u)}_{n+1/2}(u_{n+1}-u_n)^2 + \frac{1}{2}k^{(v)}_{n+1/2}(v_{n+1}-v_n)^2 + \frac{1}{2}V_n(u_n-v_n), \quad (1)$$

where u_n and v_n denote the transverse displacements from equilibrium of the two chains. No longitudinal displacements are taken into consideration. Using the Hamiltonian we obtain the equations of motion,

$$m_n\frac{d^2u_n}{dt^2} = k^{(u)}_{n+1/2}(u_{n+1}-u_n) - k^{(u)}_{n-1/2}(u_n-u_{n-1}) - \frac{1}{2}F_n(u_n-v_n), \quad (2)$$

$$m_n\frac{d^2v_n}{dt^2} = k^{(v)}_{n+1/2}(v_{n+1}-v_n) - k^{(v)}_{n-1/2}(v_n-v_{n-1}) + \frac{1}{2}F_n(u_n-v_n) \quad (3)$$

for the atoms on each chain of the double helix, where $F_n(y)=(dV_n/dy)(y)$.

Since the breathing events observed in DNA with a defective base are regular in neither frequency or duration, we wish to formulate a stochastic model of the process. We add random white noise (ξ) and damping terms to Eqs. (2) and (3) to obtain

$$m_n\frac{d^2u_n}{dt^2} = k^{(u)}_{n+1/2}(u_{n+1}-u_n) - k^{(u)}_{n-1/2}(u_n-u_{n-1}) - \frac{1}{2}F_n(u_n-v_n) - \widetilde{\eta}_n\frac{du_n}{dt} + \widetilde{\epsilon}_n\xi^u_n(t), \quad (4)$$

$$m_n\frac{d^2v_n}{dt^2} = k^{(v)}_{n+1/2}(v_{n+1}-v_n) - k^{(v)}_{n-1/2}(v_n-v_{n-1}) + \frac{1}{2}F_n(u_n-v_n) - \widetilde{\eta}_n\frac{dv_n}{dt} + \widetilde{\epsilon}_n\xi^v_n(t), \quad (5)$$

where $\widetilde{\eta}_n$ and $\widetilde{\epsilon}_n$ are the damping and noise coefficients related by the classical fluctuation-dissipation relation

$$\widetilde{\eta} = \frac{\widetilde{\epsilon}^2}{2k_B\widetilde{T}}, \quad (6)$$

where $k_B=1.38\times10^{-23}$ J K^{-1} is Boltzmann's constant and $\widetilde{T}$ is the system temperature (see [34] for details). We return to discuss this relation in detail in Sec. III C.

B. Underlying stochastic calculus

A model of a process, which is more realistic than a simple nonlinear one, is obtained by allowing some randomness of the terms or coefficients in the differential equations [35]. Øksendal analyzed equations such as

$$dX/dt = b(t,X_t) + \sigma(t,X_t)W_t, \quad (7)$$

where W_t is a stochastic process that represents the noise term, $b(t,X_t)$ is a deterministic function, while $\sigma(t,X_t)$ is the noise amplitude function. This is considered to be the small Δt limit of the discrete equation $X_{i+1}=X_i+b(t_i,X_i)\Delta t_i+\sigma(t_i,X_i)\Delta B_i$, with $X_i=X(t_i)$ being a random variable, $\Delta t_i=t_{i+1}-t_i$, and $\Delta B_i=W_{t_i}\Delta t_i$, with B_t representing the Brownian motion [35].

To solve Eq. (7) we have to choose between the Itô and Stratanovich integrals. In our case, $\sigma(t,x)$ is a constant function so the two approaches are similar, as explained in [35]. We use the Itô integral to solve our system of equations. Since we are interested in preserving the energy in the system, the stochastic differential equations also need to contain a damping parameter. Burrage *et al.* [33] analyzed the stochastic differential equation

$$\frac{d^2x}{dt^2} = f(x) - \eta s^2(x)\frac{dx}{dt} + \epsilon s(x)\xi(t), \quad (8)$$

which is based on Newton's second law of motion. This describes the position of a particle $x(t)$ via its acceleration $(d^2x/dt^2)(t)$, a deterministic force $f(x)$ related to the potential function $V(x)$ by $f(x)=-V'(x)$, a random forcing term $\epsilon s(x)\xi(t)$ with $\xi(t)$ being white noise, which has the form $\langle\xi(t)\xi(t')\rangle=\delta(t-t')$, and a damping term $\eta s^2(x)\frac{dx}{dt}$. Equation (8) can be rewritten as

$$dX_t = V_t dt, \quad (9)$$

$$dV_t = -\eta s^2(X_t)V_t dt + f(X_t)dt + \epsilon s(X_t)dW_t, \quad (10)$$

which shows that the noise term directly influences the velocity and only indirectly the distance. The function $s(x)$ is included for generality and allows the forcing and damping terms to be amplitude dependent (x dependent) while still satisfying the fluctuation-dissipation relation that the ratio of the damping coefficient to the square of the noise coefficient be inversely proportional to the temperature, that is, $\eta/\epsilon^2=1/k_BT$. Note that this relationship is maintained if one allows $\epsilon\mapsto\epsilon s(x)$ and $\eta\mapsto\eta s(x)^2$. However, in most cases, there is no need to introduce $s(x)$ in Eq. (8); hence we simply take $s(x)$ to be a constant, that is, $s(x)=1\ \forall\ x$, as can be seen in Eqs. (4) and (5).

C. Fluctuation-dissipation relations in mesoscopic systems

We are now going to consider how the form of the fluctuation-dissipation relation is changed when a system is reduced. Starting with the classic definition from Eq. (6) in the original system [Eqs. (4) and (5)], we transform these equations of motion and simplify the model by fully separating the equations. We use the substitution $x_n=u_n+v_n$ and $y_n=u_n-v_n$ and consider $k^{(u)}_{n+1/2}=k^{(v)}_{n+1/2}=k_{n+1/2}$ for all n, which has as result the following system:

$$m_n\frac{d^2x_n}{dt^2}=k_{n+1/2}(x_{n+1}-x_n)-k_{n-1/2}(x_n-x_{n-1})-\widetilde{\eta_n}\frac{dx_n}{dt}+\widetilde{\epsilon_n}[\xi^u_n(t)+\xi^v_n(t)],\tag{11}$$

$$m_n\frac{d^2y_n}{dt^2}=k_{n+1/2}(y_{n+1}-y_n)-k_{n-1/2}(y_n-y_{n-1})-F_n(y_n)-\widetilde{\eta_n}\frac{dy_n}{dt}+\widetilde{\epsilon_n}[\xi^u_n(t)-\xi^v_n(t)].\tag{12}$$

If $N(\mu,\sigma^2)$ represents a Gaussian random variable with mean μ and standard deviation σ, we have that $N(0,1)+N(0,1)=N(0,2)=\sqrt{2}N(0,1)$. Since $\xi^u_n(t)$ and $\xi^v_n(t)$ for all n, in the discrete case, represent independent Wiener processes that can be rewritten as $\sqrt{\Delta t}N(0,1)$, we obtain $\xi^u_n(t)\pm\xi^v_n(t)=\sqrt{2}\xi_n(t)$. Comparing Eqs. (11) and (12) with Eqs. (4) and (5) we note that the damping coefficients are identical ($\widetilde{\eta_n}$), but the noise coefficients are larger in Eqs. (11) and (12) than in Eqs. (4) and (5). Since the fluctuation-dissipation relation involves the noise and damping coefficients and Eqs. (11) and (12) have different noise amplitudes than (u,v) system, the (x,y) system satisfies an alternative fluctuation-dissipation relation, which will be determined later.

Observe that a breathing event is characterized by fluctuations in the distance between the two bases of each pair, which in our case is represented by the variable y_n. Hence, in what follows, we only analyze the y_n system. The mass of all bases is approximately the same, thus we may consider $m_n=m\ \forall\ n$. We analyze a particular case of this system, removing m_n from the equations by redefining the spring constant as follows: $k_{n+1/2}=mk$ for all n except for $k_{1/2}=k_{-1/2}=m\hat{k}$. Moreover, we consider $V_n(y)=\frac{1}{2}m\gamma y^2$ for $n\neq 0$ and $V_0(y)=mE_0(y)$, where E_0 is the energy function for the breathing base pair, which will be discussed later. We also have $\widetilde{\eta_n}=m\eta_n$, with $\eta_n=\eta$ for $n\neq 0$, and $\widetilde{\epsilon_n}=m\overline{\epsilon_n}$, with $\overline{\epsilon_n}=\overline{\epsilon}$ for $n\neq 0$; our system becomes

$$\frac{d^2y_n}{dt^2}=k(y_{n+1}-2y_n+y_{n-1})-\gamma y_n-\eta\frac{dy_n}{dt}+\overline{\epsilon}\sqrt{2}\xi_n(t)\quad(|n|>1),\tag{13}$$

$$\frac{d^2y_{-1}}{dt^2}=\hat{k}(y_0-y_{-1})-k(y_{-1}-y_{-2})-\gamma y_{-1}-\eta\frac{dy_{-1}}{dt}+\overline{\epsilon}\sqrt{2}\xi_{-1}(t),\tag{14}$$

$$\frac{d^2y_0}{dt^2}=\hat{k}(y_1-2y_0+y_{-1})-\frac{dE_0}{dy}(y_0)-\eta_0\frac{dy_0}{dt}+\overline{\epsilon_0}\sqrt{2}\xi_0(t),\tag{15}$$

$$\frac{d^2y_1}{dt^2}=k(y_2-y_1)-\hat{k}(y_1-y_0)-\gamma y_1-\eta\frac{dy_1}{dt}+\overline{\epsilon}\sqrt{2}\xi_1(t).\tag{16}$$

As we can see, except for $n=0$, all the interchain bounds are modeled by a linear force-displacement relationship with coefficient γ, giving a system of linear differential equations in y_n. At $n=0$ Eq. (15) gives a nonlinear force-displacement relationship. The Hamiltonian specific for the deterministic forces of our system, which generates the latter system of equations, is

$$H_y=\sum_n\left[\frac{1}{2}\left(\frac{dy_n}{dt}\right)^2+\frac{1}{2}k(y_{n+1}-y_n)^2+\frac{1}{2}\gamma y_n^2\right]+E_0(y_0)-\frac{1}{2}\gamma y_0^2+\frac{1}{2}(\hat{k}-k)[(y_1-y_0)^2+(y_0-y_{-1})^2].\tag{17}$$

Observe that we have different damping and noise coefficients at the defect site. Our computations show that we need more noise and implicitly more damping at the defect site, which implies we need different coefficients $(\overline{\epsilon_0},\eta_0)$; since $\eta=\overline{\epsilon}^2/2k_BT$, we have $\eta_0=\overline{\epsilon_0}^2/2k_BT$. The noise coefficients in this case are $\epsilon=\sqrt{2}\overline{\epsilon}$ and $\epsilon_0=\sqrt{2}\overline{\epsilon_0}$ and based on Eq. (6) we obtain that the alternative fluctuation-dissipation relation is

$$\eta=\frac{\epsilon^2}{4k_BT}\tag{18}$$

and $\eta_0=\epsilon_0^2/4k_BT$. We observe that the fluctuation-dissipation relation [Eq. (18)] for our x-y system [Eqs. (11) and (12)] has an increased noise to damping ratio of 2 over that from Eq. (6) for u-v system [Eqs. (4) and (5)]. The reason for this is that Eqs. (4) and (5) are a coupled system of $2N$ differential equations, while each of Eqs. (11) and (12) is a closed system of just N differential equations. Yet each of Eqs. (11) and (12) contains the effects of all $2N$ noise terms from Eqs. (4) and (5).

Recall that each base is considered in our model to be a separate mass point. Analyzing the four bases of a DNA duplex, we observe that the adenine (A) as well as the thymine (T) contain 32 atoms, the guanine (G) contains 33 atoms, while the cytosine (C) contains only 30 atoms. Hence we can say that on average each base contains 32 atoms and the equation of motion of each base is actually obtained from the equations of motion of the 32 atoms composing the base. Our system parameters are fitted to data obtained using AMBER, which simulates all atoms in a 12 base-pair DNA sequence solvated in water. For this reason, we consider the generalized fluctuation-dissipation relation

$$\eta = \frac{\epsilon^2}{C k_B T}, \tag{19}$$

where C is a parameter to be determined. The four bases contain different combinations of H, C, N, or O atoms. While the mass of C, N, and O are similar and that of H is negligible. Since H represent about half of the atoms of each base, we may expect $C=64$. On the other hand, the distance between the bases of a pair is the distance between the atoms from the extremities of these bases, which are linked to one or two atoms only. Thus, we have that $2<C<64$ and a precise value of C will be determined later.

The random forcing can be represented as a generalized stochastic process called *white noise* [36], in which $\xi_n(t) = dB_n(t)$ and $B_n(t)$ is continuous in time. Applying the Itô formula and discretizing, we obtain the system of equations,

$$y_n^i = y_n^{i-1} + v_n^{i-1}\Delta t_i \quad (|n|>1), \tag{20}$$

$$v_n^i = v_n^{i-1} + [k(y_{n+1}^{i-1} - 2y_n^{i-1} + y_{n-1}^{i-1}) - \gamma y_n^{i-1} - \eta v_n^{i-1}]\Delta t_i + \epsilon \Delta B_n^i \quad (|n|>1), \tag{21}$$

$$y_{-1}^i = y_{-1}^{i-1} + v_{-1}^{i-1}\Delta t_i, \tag{22}$$

$$v_{-1}^i = v_{-1}^{i-1} + [\hat{k}(y_0^{i-1} - y_{-1}^{i-1}) - k(y_{-1}^{i-1} - y_{-2}^{i-1}) - \gamma y_{-1}^{i-1} - \eta v_{-1}^{i-1}]\Delta t_i + \epsilon \Delta B_{-1}^i, \tag{23}$$

$$y_0^i = y_0^{i-1} + v_0^{i-1}\Delta t_i, \tag{24}$$

$$v_0^i = v_0^{i-1} + \left[\hat{k}(y_1^{i-1} - 2y_0^{i-1} + y_{-1}^{i-1}) - \frac{dE_0}{dy}(y_0^{i-1}) - \eta_0 v_0^{i-1}\right]\Delta t_i + \epsilon_0 \Delta B_0^i, \tag{25}$$

$$y_1^i = y_1^{i-1} + v_1^{i-1}\Delta t_i, \tag{26}$$

$$v_1^i = v_1^{i-1} + [k(y_2^{i-1} - y_1^{i-1}) - \hat{k}(y_1^{i-1} - y_0^{i-1}) - \gamma y_1^{i-1} - \eta v_1^{i-1}]\Delta t_i + \epsilon \Delta B_1^i, \tag{27}$$

which is then integrated using the implicit midpoint method [33]. Here, for each time step i and each lattice site n, ΔB_n^i is an independent normally distributed random variable with zero mean and standard deviation of $\sqrt{\Delta t_i}$.

IV. PARAMETER FITTING METHODOLOGY

The system of Eqs. (20)–(27) contains many parameters, namely, C, η, η_0, ϵ, ϵ_0, k, $\hat{k}$, γ, and the energy function $E_0(y_0)$, whose values influence the system solution. For this reason, their values have to be chosen carefully such that our model behaves in a similar manner to the experimentally observed systems and all-atom molecular dynamics (MD) simulations.

One important quantity in our system is the parameter C, which determines the ratio of noise to damping. Too much damping means no breathing events, while not enough damping allows too many breathing events to take place. Note that the interactions between the DNA atoms and the solvent surrounding it also influence the value of the constant C since the water box used in AMBER simulations slows the DNA bases. In other words, the water induces more damping into our system, which leads to a decrease in the parameter C.

One may think this value will be the same for all DNA twist angles. From structural point of view, the DNA sequence does not modify with the twist angles, however, interactions between atoms within a base and interactions between the base and its surrounding water box may depend on twist angle. This allows a breathing base to explore different volumes of space. We model this effect by varying the parameter C with twist angle. Our simulations show that $4.8 \leq C \leq 7.8$, depending on the twist angle. For each twist angle, the parameter C has been fitted to give the best fit between the SDE data and the AMBER simulations. To summarize the results presented later (Table II), the values of C used are 6.5, 5.8, 7, and 7.25 for angles of 30°, 33°, 36°, and 38°. This reduction and increase in the values of C with twist suggest that damping effects are *less* important and/or noise effects are *more* significant at more extreme twist angles. At higher twists the bases are more shielded from the effects of the aqueous environment, while at lower twists they are more exposed to it. However, the water molecules cause the random forcing as well as damping the motion of the bases, and, as shown later, other parameters are also dependent on twist and influence the breathing dynamics.

As already mentioned, the system's temperature, T, is related to η and ϵ through the fluctuation-dissipation relation. In our case, the temperature is $\tilde{T}=293$ K and hence $k_B\tilde{T} = 4.1 \times 10^{-21}$ J [5]. Note that, before introducing noise and damping in our system, we have divided each equation by the mass of a base, that is, $m=0.5098 \times 10^{-24}$ kg. Taking into account the fluctuation-dissipation relation, this implies $k_B T = k_B \tilde{T}/m$, hence $k_B T = 0.8125$ Å^2 ps^{-2}.

Most papers in the literature assume that the along-chain interactions are all identical—see, for example, [6]—and assume defects only influence the coupling between the two chains (γ and E_0). Our model enables us to test the hypothesis $k=\hat{k}$ and later results suggest $k>\hat{k}$ (see Table III), hence we treat k and $\hat{k}$ as two distinct parameters.

A. Maximum likelihood method

Using the MLE and data obtained during AMBER simulations, we determine k, $\hat{k}$, γ, ϵ, ϵ_0, and implicitly η and η_0 via the fluctuation-dissipation relation [Eq. (19)]. Note that the time step in AMBER simulations is constant, thus $\Delta t_i = \Delta t \ \forall \ i$.

Taking into account the nonlinearity of the system generated by the breathing pair, we first apply MLE method for y_1, which involves only linear terms in y_0, y_1, v_1, and y_2 to obtain the parameters k, $\hat{k}$, γ, and ϵ.

From the system [Eqs. (20)–(27)], we have $v_1^{i+1} \approx N(\mu_{i+1}, \sigma^2)$, with $\mu_i = v_1^{i-1} - [\eta v_1^{i-1} + \gamma y_1^{i-1} - k(y_2^{i-1} - y_1^{i-1}) + \hat{k}(y_1^{i-1} - y_0^{i-1})]\Delta t$ and $\sigma^2 = \epsilon^2 \Delta t$. The logarithmic likelihood is

$$l_1(\epsilon,k,\hat{k},\gamma) = \ln[L_1(\epsilon,k,\hat{k},\gamma)]$$
$$= -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(v_1^i - \mu_i)^2. \qquad (28)$$

After computing the parameter values for which the likelihood function is maximized, we compute the 95% confidence intervals for the parameters in order to determine permitted ranges for them.

Let θ be a q vector of parameters. We denote the information (a $q \times q$ matrix) by

$$I(\theta)_{ij} = \left(\mathbb{E}_x\left[-\frac{\partial^2 l_1}{\partial\theta_i\,\partial\,\theta_j}(\theta)\right]\right), \qquad (29)$$

where x is a vector of data and $1 \le i,j \le q$. Then the estimate of θ_i using MLE method is given by $\hat{\theta}_i \approx N\{\theta_i,[I^{-1}(\theta)]_{ii}\}$. Instead of I, we can use the observed information $I_{obs}(\theta) = H(\theta)$, where H is the Hessian matrix of l_1 and the variance of θ_i will then be approximately $[I_{obs}^{-1}(\theta)]_{ii}$.

Finally, since we are fitting many data sets and we might expect one to lie outside the standard confidence interval, we use the Bonferroni correction to obtain $100(1-\alpha/n)\%$ confidence interval, where n is the number of data sets tested and α is the significance level. In our case $n=8$, since we analyze eight different twist angles, and $\alpha=0.05$, hence the 99.375% confidence interval for $\widehat{\theta_i}$ become

$$[\widehat{\theta_i} - 2.5\sqrt{[I_{obs}^{-1}(\hat{\theta})]_{ii}}, \widehat{\theta_i} + 2.5\sqrt{[I_{obs}^{-1}(\hat{\theta})]_{ii}}]. \qquad (30)$$

B. Determination of interchain forces and energies

Using data from AMBER simulations, it is possible to determine the form of the force-distance relationship for the interchain separations and the associated energy function, known as potential of mean force (PMF). The standard procedure is as follows:

(i) determine the minimum min and the maximum max displacements from some reference distance between the bases of the breathing pair—for example, for a 30° twisted DNA sequence we typically have min=−0.5998 Å and max=5.1437 Å;

(ii) split the interval [min,max] into several bins of equal size s;

(iii) count the frequency f_i of each bin in the distances represented in the AMBER data;

(iv) let f_{tot} be the total number of data points available;

(v) as a first approximation, we have that for each bin i the corresponding value for PMF(y_0) is $-k_BT\ln(f_i/f_{tot})$; and

(vi) use spline interpolation to determine an expression for the potential of mean force, PMF(y_0), as illustrated in Fig. 6.

The size s of the bins influences the expression of the energy function. The number of bins N_{bin} depends on s:$N_{bin}=[(\text{max-min})/s]+1$. We have tested a wide range of bin sizes from $s=0.01$ up to $s=1$ to investigate the effect of s on PMF(y_0). In Fig. 7, we illustrate how the barrier ΔB for the base-pair breathing varies when the bin size is changed. Moreover, it shows that s should not take values below 0.2 or above 0.5 since in such cases the barrier variation with bin size is too large. When s is too large the bins are so coarse grained that the barrier is not resolved at all, leading to underestimate of ΔB, whereas when s is very small, there are so few plane paths in each bin that ΔB varies wildly with s.

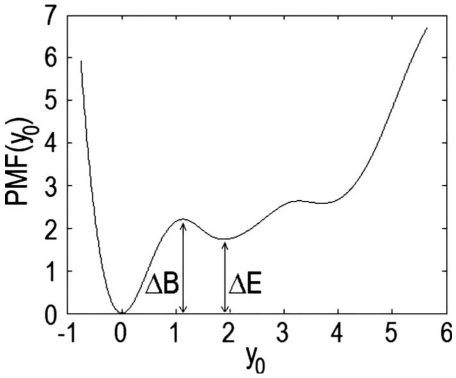

FIG. 6. Illustration of the potential of mean force, PMF(y_0) (Å^2 ps^{-2}), as a function of displacement y_0 (Å), using the bins size of s=0.5 Å, for a twist of 30°.

Note that in the full AMBER model the potential of mean force, PMF(y_0), includes the free energy of the system, that is, the potential and kinetic energies and entropy terms of the system, but in the mesoscopic model, the effects of free energy are incorporated into the potential energy, damping and forcing parameters of the stochastic differential equation model, as we shall see later.

C. Improved method for fitting energy functions to data

We apply the MLE method for the breathing pair to obtain a more accurate estimate of $E_0(y_0)$. Note that the system considered has a different noise coefficient (ϵ_0) at the defect site and consequently the damping coefficient for the breathing pair also has a different value, $\eta_0=\epsilon_0^2/Ck_BT$. Hence, we use MLE to obtain the parameters ϵ_0 and $E_0(y_0)$.

We have that $v_0^{i+1} \approx N(\mu_{i+1},\sigma^2)$, with $\mu_i = v_0^{i-1} - [\eta_0 v_0^{i-1} + (dE_0/dy)(y_0^{i-1}) - \hat{k}(y_1^{i-1} - 2y_0^{i-1} + y_{-1}^{i-1})]\Delta t$ and $\sigma^2 = \epsilon_0^2\Delta t$. The logarithmic likelihood in this case is

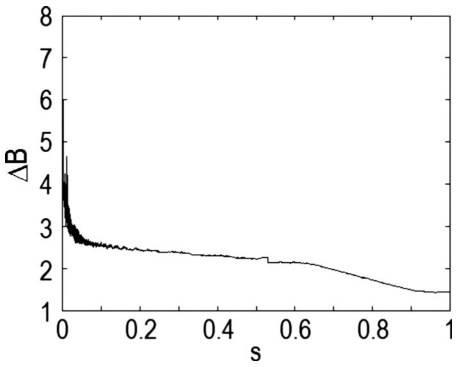

FIG. 7. Illustration of the breathing barrier ΔB (Å^2 ps^{-2}) against the bin size (s, measured in Å) (see Fig. 6 for the definition of ΔB).

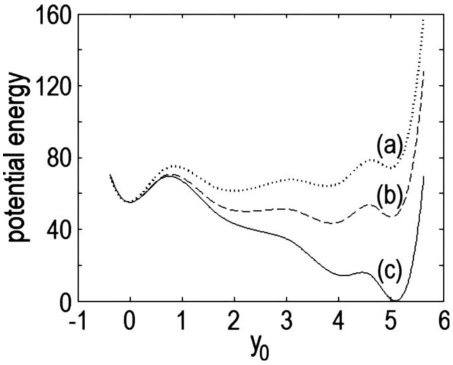

FIG. 8. Illustration of potential energy function (a) $E_0(y_0)$, (b) $E_0(y_0)+\frac{1}{2}\hat{k}y_0^2$, and (c) $E_0(y_0)+\frac{1}{2}\hat{k}y_0^2+\sqrt{k_BT}\eta_0 y_0$ (all expressed in Å^2 ps^{-2}), specific to the breathing pair of the SDE system, for a 30° undertwisted DNA.

$$l_0(E_0,\epsilon_0) = \ln[L_0(E_0,\epsilon_0)] = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(v_0^i - \mu_i)^2. \tag{31}$$

Observe that $E_0(y_0)$ is generated from a vector of pairs (x_i,y_i), with $\{x_i\}$ (an increasing array) representing the binned displacements and $\{y_i\}$ representing the value of the free energy for each bin. The final expression for $E_0(y_0)$ is obtained using a cubic spline approximation. During the maximization process only the values of $\{y_i\}$ are modified. Applying MLE method for l_0 using AMBER data for a 30° undertwisted DNA, we obtain for $E_0(y_0)$ the representation in Fig. 8(a). As we observe, this is different from Fig. 6 in a number of ways. First, Fig. 6 is obtained from a straightforward bin count of the number of time points at which the displacement falls within each interval. A fairly crude division of the interval into widths of $s=0.5$ Å is used and as noted in Fig. 7 the height of the breathing barrier is dependent on the bin width, s. As s is reduced, the accuracy will improve and Fig. 7 shows that the breathing barrier height increases. Figure 8 shows the results of a maximum likelihood estimate of the parameters followed by a calculation of the potential of mean force. We observe a significantly higher potential barrier (than in Fig. 6) since the method of calculation takes account of the order of data points in the sample data. The calculation can distinguish between a few long breathing events and many short breathing events, which is impossible when using the simpler bin-counting algorithm for estimating the PMF. As well as the potential of mean force, shown as curve (c) in Fig. 8, we plot $E_0(y_0)$, which represents the interchain potential energy (a), and the total free energy of the system [curve (b)]. We determine an expression for the total free energy of the SDE system in Sec. IV C.

D. Potential of mean force and $E_0(y_0)$

In the SDE system, the deterministic force acting on the breathing pair has several components:

(1) the along-chain force—$\hat{k}(y_1-2y_0+y_{-1})$;
(2) the interchain force—$-(dE_0/dy)(y_0)$; and
(3) the damping force—$-\eta_0 v_0$.

Using bin counts of the AMBER data, we compute the so-called potential of mean force, PMF(y_0), which includes all the deterministic forces in our system, while the MLE method considers $E_0(y_0)$ to be the energy specific to the interchain interactions. From Eq. (17) we have the total potential energy corresponding to the breathing pair being $E_0(y_0)+\frac{1}{4}\hat{k}[(y_1-y_0)^2+(y_0-y_{-1})^2]$. If we take into account the fact that the neighboring pairs do not breath and have only small deviations from equilibrium, we have that $\langle y_1\rangle=\langle y_{-1}\rangle=0$, $\langle y_{-1}^2\rangle \ll \langle y_0^2\rangle$, and $\langle y_1^2\rangle \ll \langle y_0^2\rangle$, which implies that the total potential energy is approximately $E_0(y_0)+\frac{1}{2}\hat{k}y_0^2$. Figure 8(b) shows that a graph of the total potential energy of our SDE system is closer to the potential of mean force displayed in Fig. 6, but the breathing state ($y_0=4$ Å) has a lower energy than the closed state ($y_0=0$ Å).

The damping term also contributes a deterministic force to the system since the coefficient is constant and is related only to the noise amplitude and not to the noise term itself. Consider the simple case of a moving particle subject to both deterministic and nondeterministic forces, $d^2x/dt^2=-kx-\eta\frac{dx}{dt}+\epsilon\xi(t)$. Then the total associated energy is $E(x)=K(x)+U(x)$, where $K(x)=\frac{1}{2}(\frac{dx}{dt})^2$ is the kinetic energy and $U(x)$ is the remainder. Using the fact that $d^2x/dt^2=-\partial U/\partial x$, we obtain $U(x)=\frac{1}{2}kx^2+\eta\int\frac{dx}{dt}dx$. If we interpret U as the total free energy, then the first term is clearly the potential energy and the second term is the entropic component. To calculate this latter contribution, we take $\eta \ll 1$ and $\epsilon \ll 1$ and consider that $E(x)=E_1$ fixed and then $x(t)$ is periodic, with $(\frac{dx}{dt})^2=2E_1-kx^2$. Using this value and integrating we obtain

$$\begin{aligned}\int\frac{dx}{dt}dx &= \pm\int\sqrt{2E_1-kx^2}dx\\ &= \pm\int\left[\sin^{-1}\left(x\sqrt{\frac{k}{2E_1}}\right)\right.\\ &\quad\left.+x\sqrt{\frac{k}{2E_1}\left(1-\frac{kx^2}{2E_1}\right)}\right]\frac{E_1}{\sqrt{k}}dx\\ &\approx x\sqrt{2E_1}\left(1-\frac{kx^2}{12E(x)}\right),\end{aligned} \tag{32}$$

the approximation being for small x. Since, during AMBER simulations, the energy is preserved at $E(x)\approx\frac{1}{2}k_BT$, we obtain the leading order result $U(x)=\frac{1}{2}kx^2+\eta x\sqrt{k_BT}$. Hence, the potential of mean force is given by

$$\text{PMF}(y_0) = E_0(y_0) + \tfrac{1}{2}\hat{k}y_0^2 + \sqrt{k_BT}\eta_0 y_0. \tag{33}$$

Indeed, Fig. 8(c) shows that the breathing states at $y_0=2$ Å and $y_0=4$ Å both have higher energy than the closed state $y_0=0$ Å and thus Fig. 8(c) and Eq. (33) are similar to the classic potential of mean force from Fig. 6. Hence, Eq. (33) is an approximation of the total system energy, which can be computed only based on our system parameters values.

V. SDE RESULTS

We are interested in understanding the MD simulations on twist-dependent breathing in DNA strands with defective

TABLE I. Time spent breathing for each angle analyzed using different numbers of data points.

Twist angle (deg)	14 ns	13 ns	12 ns	11 ns	10 ns
30	26.9500%	26.8769%	25.1583%	26.7000%	24.7600%
32	45.1143%	48.5000%	48.4000%	50.7909%	46.3300%
33	23.8786%	25.4385%	27.3750%	29.5273%	27.5400%
34	21.7571%	23.1692%	24.9583%	21.0636%	23.0500%
35	26.2500%	24.3538%	25.6500%	27.5545%	28.2600%
36	40.1357%	38.0538%	37.9833%	35.7455%	37.9500%
38	64.1143%	65.1462%	69.0917%	70.2636%	70.9000%
40	57.0429%	55.9000%	58.0250%	63.2364%	66.9900%

bases. Since the interaction of DNA with other macromolecules may locally alter its twist and hence the frequency of breathing, we wish to understand how changing the twist influences the proportion of time spent in the open breathing state and typical length of such breathing events. Normally twisted DNA has an angle of $35°-36°$ per base pair. It is interesting to note that undertwisting causes a decrease in length of breathing events, while overtwisting causes an increase in the proportion of time spent breathing, although the mechanisms for the two cases appear similar.

Using AMBER, we have simulated 20 ns of data for each twist angle analyzed. We ignore the first 5 ns of each simulation since the data show an unrepresentative initial transient. Table I shows that the time spent breathing is different for different portions of data analyzed, but the way in which it varies with the twist angle is preserved no matter which sample interval is used.

From the 20 ns simulations, which provide data about each 1 ps, we calculate how long the A-F pair is in a breathing state (as a percentage). We then choose a shorter simulation, of 1 ns, for example, in which the same percentage of time is spent breathing. For this later run, we keep the position and velocity every 2 fs in order to obtain accurate parameters values using MLE method. It is impossible to store data every 2 fs for 20 ns since such simulations would require more than 10 weeks and about 8000 Gbytes of storage capacity.

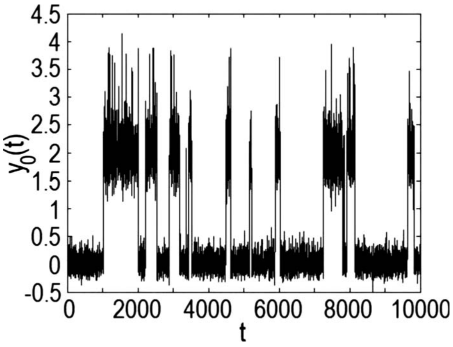

FIG. 9. Illustration of the variation of the distance (Å) between the bases of the breathing pair for 10 ns obtained using the proposed model for a 30° undertwisted DNA sequence. The parameter values are $C=6.5$, $\epsilon=3.4074$, $\epsilon_0=5.6285$, $k=10.6536$, $\hat{k}=3.6851$, $\gamma=120.0904$, $\eta=\epsilon^2/Ck_BT$, and $\eta_0=\epsilon_0^2/Ck_BT$, while for E_0 the expression from Fig. 8 was used.

For each angle, we compute the parameter values from simulation data corresponding to the smallest and the largest proportions of time spent breathing. In this way we obtain two confidence intervals for each parameter, which we combine to give the final intervals for our parameters.

A. Kinetics of the mesoscopic model

We simulate our SDE system using the implicit midpoint method. A comparison with AMBER simulations shows that the two approaches have similar results, which implies that the system definition and parameters fitting methodology are consistent.

We simulated the SDE system for a 30° undertwisted DNA (Fig. 9) and obtained a result similar to the one from AMBER (Fig. 1). Some differences may be observed: the AMBER data suggest that the oscillation interval is between -0.3 and 5.7 Å (a 6 Å range), while in our case we have oscillations between -0.3 and 4.2 Å (a 4.5 Å range). One explanation for this reduction is the mass of the particles appearing in our system, which was eliminated from our equations by redefining the parameters. We considered that an entire base is moving, while in reality just a part of it is moving completely, while the rest remains more or less in the initial position. Moreover, our system contains only one degree of

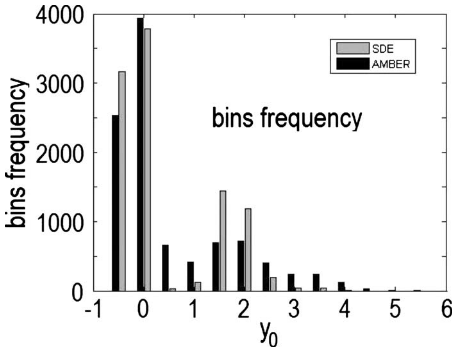

FIG. 10. Illustration of the occupation of different y_0 positions (Å) for the breathing pair, obtained from the SDE and AMBER simulations using a bin size of $s=0.5$, for a 30° undertwisted DNA sequence.

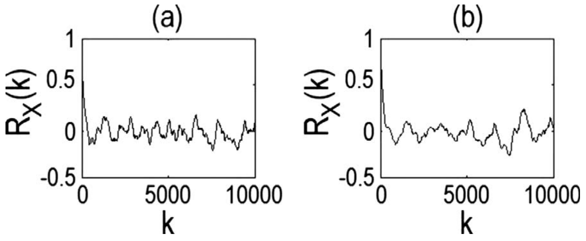

FIG. 11. Illustration of the autocorrelation function of the positions of the breathing pair, y_0 (Å), obtained from the SDE and AMBER simulations for a 30° undertwisted DNA sequence. The horizontal axis is in ps.

freedom for each base pair, while AMBER uses on average 90 degrees of freedom per base pair. The water box also influences the DNA dynamic during a simulation.

Figure 10 contains a comparison between the AMBER and SDE systems in terms of the binned frequency data over the 10 ns simulations. In the closed state (at $y_0=0$ Å), the residence time is similar, however, we observe a reduction in the number of data points at the breathing barrier $\Delta B \approx 1$ Å (see Fig. 6 for definition) and an increased number of points for the bins corresponding to the breathing state at $y_0=2$ Å. This is counterbalanced by the residence time at $y_0=4$ Å, which is reduced in the SDE simulation compared to AMBER and hence the total time spent breathing is similar, that is, 28.71% of the simulation time. Note that graphs such as Fig. 10 depend on the width of bins chosen; using wider bins would increase the accuracy of the results on the vertical axis but result in a lower resolution of the detail of the closed and open states, which is a lower resolution on the horizontal axis. Similarly, it was noted in Sec. IV B and Fig. 7 that the height of the breather barrier is dependent on the width of the bins used since the small time spent near the barrier means that there is a relatively low number of counts there and the relative errors are larger.

For a discrete random variable X of length n, mean μ_X, and standard deviation σ_X, the autocorrelation function is defined to be

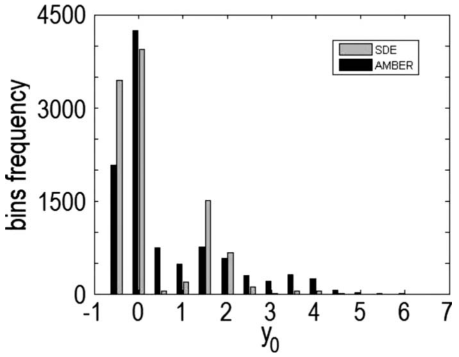

FIG. 13. Illustration of the occupation of different y_0 positions (Å) for the breathing pair, obtained from the SDE and AMBER simulations using a bin size of $s=0.5$, for a 33° undertwisted DNA sequence.

$$R_X(k) = \frac{1}{(n-k)\sigma_X}\sum_{i=1}^{n-k}(X_i-\mu_X)(X_{i+k}-\mu_X), \quad \forall \quad 0 \le k < n. \tag{34}$$

In Fig. 11 the autocorrelation function for the DNA sequence with a twist of 30° is plotted. On the left (a) is the result from the AMBER simulation and graph (b) on the right shows the corresponding figure from our SDE simulation. The two graphs show that the two systems lose memory of their initial conditions on the same time scale and have similar features in their autocorrelation functions over longer time separations.

The SDE simulation presented in Fig. 12 emphasizes that the results obtained using our SDE model agree with the MD simulations from Fig. 2 in length and frequency of breathing events for a 33° undertwisted DNA sequence. In both closed and open states, the fluctuations are slightly smaller in the SDE model than in the full MD AMBER simulation. This can be attributed to the reduction in the number of degrees of freedom as one moves from an all-atom simulation to a mesoscopic model.

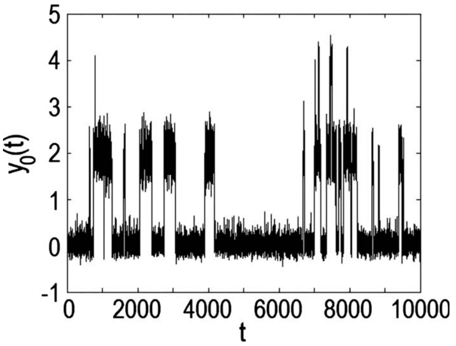

FIG. 12. Illustration of the variation of the distance (Å) between the bases of the breathing pair for 10 ns, obtained using the proposed model for a 33° undertwisted DNA sequence. The parameter values are $C=5.8$, $\epsilon=3.3429$, $\epsilon_0=5.3214$, $k=9.5374$, $\hat{k}=2.8261$, $\gamma=135.5951$, $\eta=\epsilon^2/Ck_BT$, and $\eta_0=\epsilon_0^2/Ck_BT$, while for E_0 the expression from Fig. 19(a) was used.

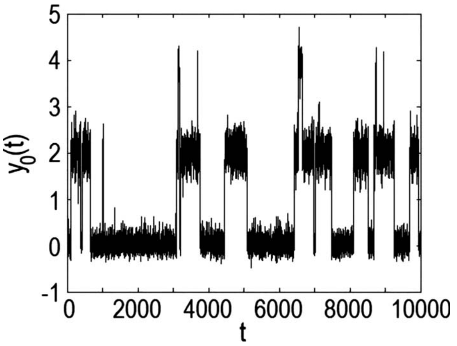

FIG. 14. Illustration of the variation of the distance (Å) between the bases of the breathing pair for 10 ns, obtained using the proposed model for a 36° twisted DNA sequence. The parameter values are $C=7$, $\epsilon=3.3499$, $\epsilon_0=5.9238$, $k=7.6577$, $\hat{k}=1.4307$, $\gamma=165.4327$, $\eta=\epsilon^2/Ck_BT$, and $\eta_0=\epsilon_0^2/Ck_BT$, while for E_0 the expression from Fig. 19(a) was used.

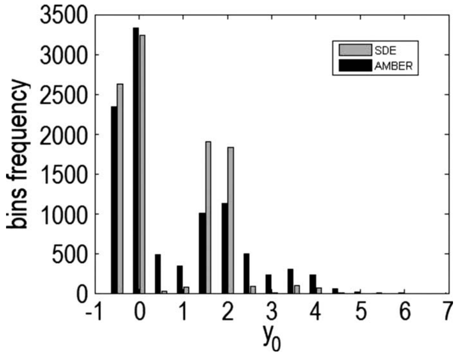

FIG. 15. Illustration of the occupation of different y_0 positions (Å) for the breathing pair, obtained from the SDE and AMBER simulations using a bin size of $s=0.5$, for a 36° twisted DNA sequence.

Indeed, Fig. 13 shows that the residence time in both open and closed states is larger in the SDE simulation than in AMBER, but the number of barrier crossings is higher in the case of AMBER simulation. This is due to the SDE simulation not exhibiting some of the very short breathing events observed in the AMBER simulation. However, overall the time spent breathing during the SDE simulation (25.74%) agrees well with the data obtained using AMBER (see Table I).

Analyzing the SDE simulation from Fig. 14, presenting a normally twisted DNA, we again observe differences in the range of values compared with the AMBER simulation from Fig. 3. Moreover, the SDE simulation is regular, the three breathing states being well defined, while in the AMBER simulation the degree of randomness is larger. On the other hand, the breathing length and frequency are approximately the same in both SDE and AMBER simulations.

Comparing the results presented in Fig. 15 with the undertwisted case (Fig. 10), we observe an increased number of data points in the second breathing state at $y_0=4$ Å. This increase occurs in both the AMBER and the SDE systems, though in all twist angles, there AMBER shows more time in the second breathing state than the SDE system. Even though the SDE simulation has a larger amount of data around the first breathing state, $y_0=2$ Å, the percentage of time spent in a breathing state is the same in both AMBER and SDE simulations, namely, 40.95%.

In Fig. 16 we compare the autocorrelation from the AMBER simulation of a 36° twisted DNA sequence [left panel, denoted (a)] to that from the SDE system [on the right, denoted (b)]. As in Fig. 11 the agreement is good, however, both show a slightly longer memory for a twist of 36° than that for 30°.

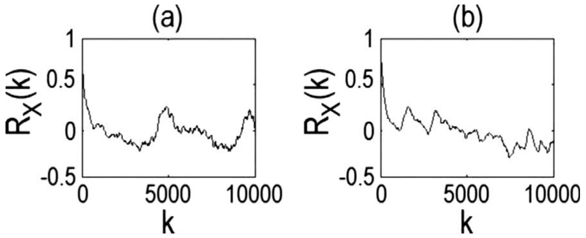

FIG. 16. Illustration of the autocorrelation function of the positions of the breathing pair, y_0 (Å), obtained from the SDE and AMBER simulations for a 36° undertwisted DNA sequence. The horizontal axis shows zero to 10 ns.

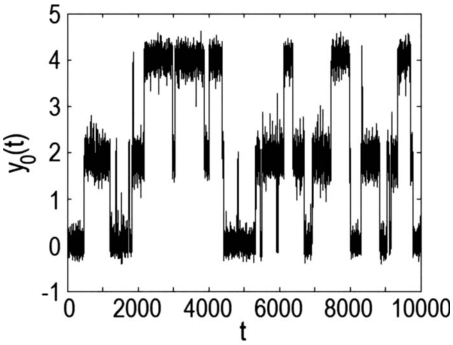

FIG. 17. Illustration of the variation of the distance (Å) between the bases of the breathing pair for 10 ns, obtained using the proposed model for a 38° overtwisted DNA sequence. The parameter values are $C=7.25$, $\epsilon=3.3511$, $\epsilon_0=6.8702$, $k=8.1438$, $\hat{k}=2.1462$, $\gamma=139.0797$, $\eta=\epsilon^2/Ck_BT$, and $\eta_0=\epsilon_0^2/Ck_BT$, while for E_0 the expression from Fig. 19(a) was used.

Finally, the SDE simulation for a 38° overtwisted DNA sequence is presented in Fig. 17. This simulation is also close to the MD simulation since 70.22% of the time is spent breathing compared to the average of 67.91% from AMBER case (see Table I). This simulation also confirms the regularity of the SDE simulations and shows that the volume of space explored by the breathing pair is indeed larger for an overtwisted DNA sequence than in the undertwisted case.

For 38° of twist, Fig. 18 shows that more time is spent in the two breathing states at $y_0=2$ Å and $y_0=4$ Å in the SDE simulation than in the AMBER data (Fig. 4). Less data points are observed near the breathing barriers at $y_0=1$ Å and $y_0=3$ Å. Even though this implies a small reduction in breathing frequency, that is, 9 breathing events in SDE simulation instead of 12 as in AMBER, the general DNA behavior is preserved. Compared to the undertwisted and normally twisted DNA sequence, in both AMBER and SDE systems we have a high residence time in the second breathing state ($y_0=4$ Å).

In conclusion, our SDE system reproduces the DNA behavior observed in all-atom MD simulation and preserves the

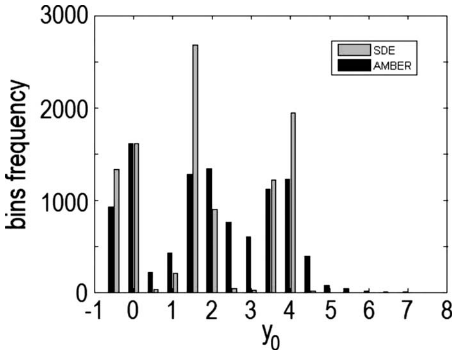

FIG. 18. Illustration of the occupation of different y_0 positions (Å) for the breathing pair, obtained from the SDE and AMBER simulations using a bin size of $s=0.5$, for a 38° overtwisted DNA sequence.

TABLE II. Parameter C values.

Twist angle	30°	32°	33°	34°	35°	36°	38°	40°
C	6.5	6	5.8	5.6	4.8	7	7.25	7.8

twist-dependent breathing properties, such as length and frequency. However, the most important features of our model are given by the system parameters. Their values allow us to understand how the along-chain and intrachain interactions vary with twist angle, as well as the twist dependency of the interactions between the DNA molecule and the surrounding water box.

B. Analysis of parameters values

As already mentioned, the value of C from the fluctuation-dissipation relation differs for each twist angle. The values used for this parameter are presented in Table II.

Fitting the parameters via MLE for all the twist angles and selecting a value inside the intervals obtained for each parameter, we end with the values listed in Table III.

As can be seen, the along-chain bonds k and $\hat{k}$ become weaker as the twist angle is increased from 30°, while the interchain bond γ becomes stronger. Once the DNA becomes overtwisted (twist angle greater than 36°), the along-chain bonds become stronger and the interchain bonds decrease. From 36° upward we see a 20.19% decrease in γ and 90% increase in $\hat{k}$.

The noise coefficient ϵ is almost constant, varying by only 0.2%, while for the A-F pair we observe small oscillations, of 15.98%, in the noise coefficient ϵ_0.

Note that the parameter γ is fitted to AMBER data for an A-T base pair ($n=1$). The bases of such a pair are linked by two hydrogen bonds and the bases of a C-G pair are linked by three hydrogen bonds, but our model does not take into account which type of base pairs our DNA sequence contains. Supposing that each hydrogen bond has equal contribution to the interactions between the bases of a pair, we use for our simulations an average value between the two case, which in our case is $5\gamma/4$.

For each angle, we obtain a different expression for the energy function $E_0(y_0)$. Figure 19(a) represents $E_0(y_0)$ for the 33°, 36°, and 38° twist angles. Some of the differences between the expressions for $E_0(y_0)$ for several angles are presented in Table IV. Here ΔB is the height of the barrier from the closed state and ΔE is the energy difference between the breathing (open) and normal (closed) states. Hence, the energy barrier from open to closed state is $\Delta B - \Delta E$. The energy differences ΔB and ΔE control the frequency and the length of breathing events, respectively, and vary with twist angle.

TABLE III. Parameter values obtained using MLE, k, $\hat{k}$, and γ, are measured in ps^{-2}, while ϵ and ϵ_0 are measured in Å ps$^{-3/2}$.

Twist angle (deg)	k	$\hat{k}$	γ	ϵ	ϵ_0
30	10.6536	3.6851	120.0904	3.4074	5.6285
32	9.5585	3.2132	131.0919	3.3585	5.9770
33	9.5374	2.8261	135.5951	3.3429	5.3214
34	9.2678	2.4625	145.6987	3.3225	5.4843
35	8.1819	1.8256	149.5683	3.3471	5.6744
36	7.6577	1.4307	165.4327	3.3499	5.9238
38	8.1438	2.1462	139.0797	3.3511	6.8702
40	19.5258	2.6741	132.0332	3.3550	6.2357

If, for an undertwisted DNA sequence, ΔE is negative, as seen in Fig. 19(a) and Table IV, for the typical twist of 36° its value is close to zero [see Fig. 19(a)]. Taking into account that $\eta_0 = 6.0089$ for the 33° twist angle, while for the 36° case we have $\eta_0 = 6.1699$, the damping contribution to the potential of mean force is the same. Hence, the stacking interaction parameter $\hat{k}$ controls the length of the breathing events.

Indeed, the value of $\hat{k}$ has the most dramatic variation: it decreases with twist angle until the typical twist angle (36°) is reached and increases with overtwist. A higher value of $\hat{k}$ means higher energy in the open state and less time spent breathing. The energy function $E_0(y_0)$ also controls the length of the breathing events through the value of ΔE, which decreases—compare the expressions from Fig. 19(a)—when the time spent breathing is larger (see Table I) or to compensate a higher value of $\hat{k}$.

Finally, the approximations of the potential of mean force for 33°, 36°, and 38° twist angles, presented in Fig. 19(b), show that the damping and especially the harmonic interchain contribution to the total system energy define the values for the displacements from equilibrium for the A-F pair between −0.3 and 5 Å. This analysis shows that breathing can be viewed as competition between the along-chain elastic energies, the interchain binding energy and the damping induced by the solvent, which slows the DNA atoms and changes the dynamics of our DNA molecule. The variation of parameters with twist angle suggests that is more likely to observe breathing in an overtwisted DNA sequence than in an undertwisted one.

The variation of breathing events is interesting: at 34°–35° breathing events are relatively rare, while for the other undertwisted DNA plasmids we observe an increase in

TABLE IV. Values of ΔB and ΔE (both measured in Å^2 ps^{-2}) corresponding to $E_0(y_0)$ (see Fig. 6 for their definition).

Twist angle	30°	32°	33°	34°	35°	36°	38°	40°
ΔB	13.9853	8.3315	12.4900	12.8309	7.6100	19.2640	13.6796	14.0867
ΔE	−11.5855	−11.0732	−5.3957	−4.0070	−1.8403	0.6502	−7.1785	−9.7745

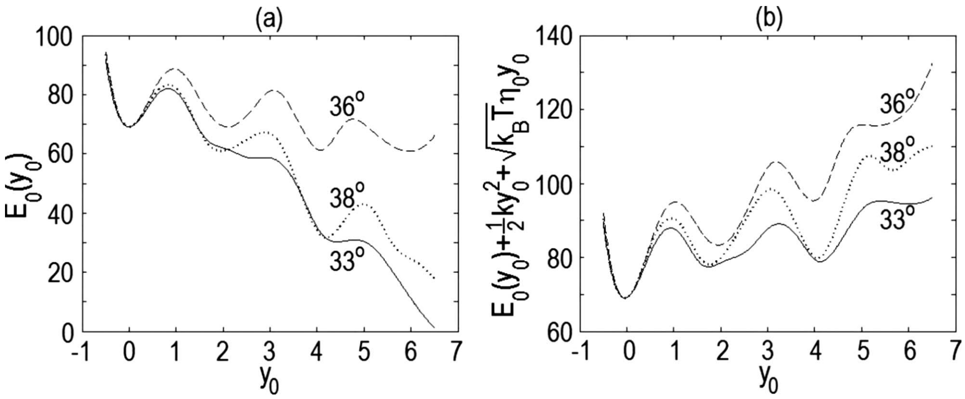

FIG. 19. Illustration of (a) interchain potential $E_0(y_0)$ (Å^2 ps^{-2}) and (b) potential of mean force (Å^2 s^{-2}) plotted against A-F bonds length (Å), obtained after fitting parameters for a 33° undertwisted DNA.

their frequency due to a reduction in the energy difference ΔE (see Table IV). For overtwisted plasmids there is again a reduction in ΔE, and also a decrease in ΔB and an increase in C (see Table II), thus less damping. This leads to a larger residence time in the breathing state, hence longer breathing events.

VI. CONCLUSIONS

This paper introduces a stochastic differential equation model for a DNA duplex useful for simulating short time scale breathing events at a defect. After presenting the nonlinear model, which incorporates noise and damping terms, we use the maximum likelihood estimation (MLE) method to determine model parameters and potential energy functions using data from the MD-simulation package AMBER. Although we observe a slight reduction in the amplitude of fluctuations in the reduced SDE model when compared with AMBER data, the time spent breathing, as well as the length and frequency of breathing events, is similar.

This paper also discusses the role of the fluctuation-dissipation relations in the derivation of reduced mesoscopic models. The model emphasizes the difference between the potential of mean force and the various potential energies in our system by showing the importance of the damping term in preserving the system energy and the way in which the along-chain interactions influence the length of a breathing event. Finally, parameter values for twist angles between 30° and 40° are presented, as well as a comparison between the dynamics of solutions of the proposed method and the data obtained using AMBER, which underlines the capability of the SDE system to simulate with accuracy breathing events.

Previously, it has been thought that breathing events were due to inhomogeneities in the interstrand interactions. However, our results also show that there is, in addition, a significant change in along-chain interactions, which contributes to the breathing. Moreover, the DNA helical twist is also important for the breathing events, length and frequency, possibly due to a change in interactions between the DNA molecule and the surrounding solvent.

In an accompanying paper [37] we give a more detailed interpretation of the results and the insights into DNA structure and dynamics which they yield.

ACKNOWLEDGMENTS

C.-I.D. was funded by the European Union as part of MMBNOTT—an Early Training Research Programme in Mathematical Medicine and Biology. We would also like to thank the referees for their helpful comments.

[1] E. Cubero, E. C. Sherer, F. J. Luque, M. Orozco, and C. A. Laughton, J. Am. Chem. Soc. **121**, 8653 (1999).
[2] J. D. Watson and F. H. C. Crick, Nature (London) **171**, 737 (1953).
[3] L. V. Yakushevich, *Nonlinear Physics of DNA* (Wiley, Chichester, 1998).
[4] L. V. Yakushevich, Physica D **79**, 77 (1994).
[5] J. A. D. Wattis, S. A. Harris, C. R. Grindon, and C. A. Laughton, Phys. Rev. E **63**, 061903 (2001).
[6] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[7] S. Takeno and S. Homma, Prog. Theor. Phys. **70**, 308 (1983).
[8] Y. Kafri, D. Mukamel, and L. Peliti, Eur. Phys. J. B **27**, 135 (2002).
[9] G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, Phys. Rev. Lett. **90**, 138101 (2003).
[10] T. Ambjörnsson, S. K. Banik, O. Krichevsky, and R. Metzler, Phys. Rev. Lett. **97**, 128105 (2006).
[11] M. Salerno, Phys. Rev. A **44**, 5292 (1991).
[12] V. Muto, P. S. Lomdahl, and P. L. Christiansen, Phys. Rev. A **42**, 7452 (1990).
[13] L. L. Van Zandt, Phys. Rev. A **40**, 6134 (1989).
[14] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, 684 (1993).
[15] J.-L. Ting and M. Peyrard, Phys. Rev. E **53**, 1011 (1996).
[16] M. Peyrard and J. Farago, Physica A **288**, 199 (2000).
[17] M. Barbi, S. Cocco, and M. Peyrard, Phys. Lett. A **253**, 358 (1999).
[18] G. Gaeta and L. Venier, Phys. Rev. E **78**, 011901 (2008).
[19] T. Ambjörnsson and R. Metzler, Phys. Rev. E **72**, 030901(R) (2005).
[20] R. Metzler and T. Ambjörnsson, J. Biol. Phys. **31**, 339 (2005).
[21] A. Hanke and R. Metzler, J. Phys. A **36**, L473 (2003).
[22] S. K. Banik, T. Ambjörnsson, and R. Metzler, Europhys. Lett.

71, 852 (2005).

[23] E. Lennholm and M. Hornquist, Physica D **177**, 233 (2003).

[24] N. R. Quintero, A. Sánchez, and F. G. Mertens, Eur. Phys. J. B **16**, 361 (2000).

[25] M. Peyrard, S. C. López, and G. James, Nonlinearity **21**, T91 (2008).

[26] L.-Y. Zhang, H. Sun, and J.-T. Lin, Phys. Lett. A **259**, 71 (1999).

[27] G. Kalosakas, K. Ö. Rasmussen, and A. R. Bishop, Synth. Met. **141**, 93 (2004).

[28] M. Peyrard, S. C. López, and D. Angelov, Eur. Phys. J. Spec. Top. **147**, 173 (2007).

[29] K. M. Guckian, T. R. Krugh, and E. T. Kool, Nat. Struct. Mol. Biol. **5**, 954 (1998).

[30] J. A. D. Wattis, Philos. Trans. R. Soc. London, Ser. A **362**, 1461 (2004).

[31] T. A. Evans and K. R. Seddon, Chem. Commun. (Cambridge) 1997, 2023.

[32] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross, and P. A. Kollman, AMBER 9, University of California, San Francisco, 2006.

[33] K. Burrage, I. Lenane, and G. Lythe, SIAM J. Sci. Comput. (USA) **29**, 245 (2007).

[34] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 3rd ed. (Springer, Berlin, 2004).

[35] B. Øksendal, *Stochastic Differential Equations*, 6th ed. (Springer, New York, 2005).

[36] T. Hida, *Brownian Motion* (Springer-Verlag, New York, 1980).

[37] C. I. Duduială, J. A. D. Wattis, I. L. Dryden, and C. A. Laughton, *Variation of DNA Parameters with Twist for a DNA Sequence with a Defect* (unpublished); C. I. Duduială, Ph.D. thesis, University of Nottingham, 2009 (unpublished); etheses.nottingham.ac.uk