Improving the mesoscopic modeling of DNA denaturation dynamics

Marta Marty-Roda[1,2]**, Oda Dahlen**[3]**, Titus S. van Erp**[3,4]**, and Santiago Cuesta-López**[1,2]

[1]International Research Center in Critical Raw Materials-ICCRAM, University of Burgos, Plaza Misael Bañuelos s/n, 09001 Burgos, Spain

[2]Advanced Materials, Nuclear Technology and Applied Bio/Nanotechnology. Consolidated Research Unit UIC-154. Castilla y Leon. Spain.University of Burgos. Hospital del Rey s/n, 09001 Burgos, Spain.

[3]Department of Chemistry Norwegian University of Science and Technology (NTNU) Høgskoleringen 5, 7491 Trondheim, Norway

[4]Center for Molecular Modeling (CMM), Ghent University, Technologiepark 903, B9000 Ghent, Belgium

E-mail: `titus.van.erp@ntnu.no`,[*]`scuesta@ubu.es`

Abstract. Although previously developed mesoscopic DNA models have successfully reproduced thermodynamic denaturation data, recent studies show that these overestimate the rate of denaturation by orders of magnitude. Using adapted Peyrard-Bishop-Dauxois (PBD) models, we calculated the denaturation rates of several DNA hairpins and made comparison with experimental data. We show that the addition of a barrier at the onsite potential of the PBD model gives a more accurate description of the unzipping dynamics of short DNA sequences. The new models provide a refined theoretical insight on the dynamical mechanisms of unzipping which can have implications for the understanding of transcription and replication processes. Still, this class of adapted PBD models seems to have a fundamental limitation which implies that it is not possible to get agreement with available experimental results on the dynamics of long DNA sequences and at the same time maintain the good agreement regarding its thermodynamics. The reason for this is that the denaturation rate of long DNA chains is not dramatically lowered by the additional barrier as the base-pairs that open are more likely to remain open, facilitating the opening of the full DNA molecule. Some care has to be taken since experimental techniques capable to study denaturation rates of long sequences seem not in agreement with other experimental data on short DNA sequences. Further research, both theoretical and experimental, is therefore needed to resolve these inconsistencies, which will be a starting point for new minimalistic models that are able to describe both thermodynamics as well as dynamics at a predictive level.

1. Introduction

Understanding the physics of deoxyribonucleic acid (DNA) is a challenge since it is a high-dimensional system with a complex structure. All-atom simulations are not only computational intensive, but also difficult to analyze. Simple models of DNA can, therefore, help to identify the essential features underlying its remarkable properties [1]. Despite the fact that a half century has passed since the discovery of the structure of the DNA double helix [2], major questions still remain regarding its thermodynamic behavior and stability. Knowledge of the DNA structure alone is generally not sufficient for understanding the biological processes [3, 4]. In addition to the structural information, it is necessary to know how the structure affects the equilibrium properties and the dynamics of the DNA molecule. One of the phenomena that can be studied is the local opening of DNA, which is an intriguing phenomenon from a statistical-physics point of view, but is also essential for its biological function. For instance, the unwinding of the double helix and the formation of specific DNA bubbles are prerequisites for both DNA replication and gene transcription [5].

In addition, understanding how base-pairs separate under thermal fluctuations, i.e thermal denaturation of DNA, is of importance to many biotechnology applications, including the design of probes for primers, in polymerase chain reaction, or molecular beacons (DNA hairpins). Besides, DNA denaturation is becoming important for nanotechnology [6–9] as DNA is now used for its self-assembly properties to create nano-devices or to design molecular memories.

In the last decades, significant progress have been made in the development of Ising-type models [10] which describe the state of a base pair by either 0 or 1 depending on whether it is closed or open, respectively. Examples are nearest-neighbor models [11], Poland-Scheraga models [12], and Benham's model [13, 14]. Other integer-type models are the zipper models [15, 16] that assume that the DNA strand can only open from one side (neglecting the effect bubbles). In these models, the state of the molecule is determined by a single integer variable: the number of open base-pairs. Models with a more continuous character are the bead-spring rotating strands [17] and the Peyrard Bishop Dauxois (PBD) model [18], that is used in the article [18]. In the latter, each base-pair is described by a single variable describing the stretch between bases. It has been shown that the PBD models can reproduce the equilibrium and thermodynamic properties of DNA reasonably well [19–21].

For a long time, the study of DNA dynamics has been given very little attention. One reason for this is the experimental difficulty to measure kinetic data such as rates of opening or unwinding. As a result, there has been a lack of accurate experimental data about DNA dynamics to which theoretical models can be compared. Another reason is that most relevant dynamical processes, which can be compared with experiments, require very long timescales. Even if highly coarse-grained models, like PBD, can straightforwardly be used in molecular dynamics and are considerably faster than all-atom models, the accessible timescale of the simulations is generally not long enough

to observe the specific biological transitions such as DNA denaturation. As a result, theoretical studies trying to analyze DNA dynamics have been focusing on biological phenomena with very short life-times, such as the opening of individual base-pairs [21] instead of the opening of large DNA bubbles or full molecules.

However, utilizing the fact that the PBD model has only first-neighbor interactions, the rate constants of DNA denaturation can be computed very accurately using the approach of Refs. [6, 22, 23] that is used in this work and further discussed in Sec. 3 The same algorithmic technique has been used for the computation of denaturation rates of double stranded DNA [6, 22] and DNA hairpins [23] based on the standard PBD model. Since the latter process is easier accessible with experimental techniques, there is significantly more experimental data available on the denaturation rate of DNA hairpins compared to double stranded DNA.

DNA hairpins are made of short single strands of DNA with terminal regions consisting of complementary base sequences. As a result, the two end-regions can self-assemble in a short DNA double helix, called the stem, while the remaining central part of the strand forms a loop. DNA hairpins have an important biological function since they take part in many RNA configurations, are involved in key biological process like the regulation of gene expression, DNA recombination, and regulation of the transcription by binding proteins [24–26].

Several experimental studies have increased our knowledge about the kinetics of the hairpin transitions by measuring the characteristic opening and closing rates [27–35]. Regarding theoretical studies, Hanne *et* al. [28] compared experimental single-molecule measurements of the DNA hairpin opening rate under mechanical tension with theoretical modeling using the PBD model. Denaturation rates of short hairpins were computed in a qualitative manner using an artificial Monte Carlo type of dynamics. This approach provided trends on the behavior of relative rates as function unzipping force and other parameters like temperature and sequence content. Dahlen and van Erp [23] were the first to report quantitative computations of hairpin denaturation rates using the PBD-class specific reactive flux implementation described in Sec. 3. Their results show that the PBD model, despite giving reasonable results regarding thermodynamic properties, overestimate rates of unzipping by many orders of magnitude.

In this work, we will therefore examine the dynamical properties regarding hairpin denaturation using a recent adaptation of the PBD model [21] which has shown to give better agreement regarding base-pair lifetimes. In addition to the model itself, we examined some variations of the model to explore the effect of some model parameters on the behavior of denaturation rates as function of temperature and chain length.

2. Adapted PBD model

The PBD model describes the double stranded DNA molecule as a connected chain of one-dimensional particles and condenses all the atomic coordinates of a base-pair into a single number y, which describes the stretching of the bonds between the two

bases [3, 18]. Despite its simplicity, the model successfully describes the sharp melting transition of long DNA homopolymers [18], provides reasonable statistics on bubbles in DNA sequences [36], and on the persistence length and DNA flexibility [37]. However, the average time during which a base-pair stays closed between two opening fluctuations, and the time during which it stays open after such a fluctuation are orders of magnitude too small [21] in comparison to experimental values inferred from proton-deuterium exchange experiments [38].

A way to increase the life-times of the open state and the closed state of individual base-pairs is to add a barrier to the onsite potential. Several studies have argued that such an adaptation gives a better physical description of DNA and have proposed to replace the original onsite Morse potential in the PBD model by a potential that has a true barrier in between the open and the closed state [21, 39, 40]. The suggested adaptations [21, 39, 40] of the potential function differ with respect to their mathematical formulations, but are rather similar in nature. Besides making it more difficult for the base-pair to open, the modification also introduces a obstacle for the closing process. Whereas in the original PBD model this transition is barrier-free, in the new models an energy threshold needs to be overcome. This closing barrier is a result of the hydrogen bonds that the bases make with the solvent that need to be broken to make the transition [39, 41]. Besides enthalpic contributions, the closing barrier also reflects the loss of degrees of freedom when the base-pair gets packed into the helical structure [21], which has been quantified in all-atom free energy calculations [42]. As mesoscopic models aim to include entropy by effective potentials of mean force, the decrease in entropy is manifested by a higher potential value in the coarse-grained description. This value goes rapidly down when bringing the bases closer together since the loss of entropy and solvent interactions is more than compensated by the favorable energetic interaction between the nearby opposite bases.

The inclusion of the barrier provides base-pair life-times that are considerably closer to the experimental values [21]. Calculations on the melting transition of the A-form of DNA using the adaptation of the PBD model [21] has also given a good agreement with with calorimetry and neutron scattering experiments [43]. Yet, it is an open question whether this improved model, possibly with slight variations in model parameters, is able to give a correction description of the denaturation rate of DNA polymers and hairpins. The denaturation process or the opening of the full DNA chain requires many base-pair opening transitions. The opening of a single base-pair or the simultaneous opening of a few consecutive base-pairs is generally not followed-up by the opening of the full molecule. As long as the majority of base-pairs are in a closed state, it is much more likely that the open base-pairs will be pulled back into the stack. The effect of the barrier on the denaturation rate is, therefore, non-trivial and has not been studied before.

The Hamiltonian for DNA chain with N base-pairs in the adapted PBD model is

the following:

$$H = \sum_{n=1}^{N} \frac{p_n^2}{2m} + W(y_n, y_{n-1}) + V(y_n), \text{ with } p_n = m\frac{dy_n}{dt} \tag{1}$$

where $W(y_n, y_{n-1})$ for $n > 1$ is an anharmonic potential which describes the interaction between adjacent bases along the DNA molecule. $V(y)$ is the Morse potential with its modification, describing the repulsive interaction due to the hydrogen bonds between the two bases of a base-pair on opposite strands

$$W(y_n, y_{n-1}) = \frac{1}{2}K(1 + \rho e^{-\kappa(y_n + y_{n-1})})(y_n - y_{n-1})^2 \tag{2}$$

$$V(y) = D(e^{-\alpha y} - 1)^2 + \theta(y)\frac{by^3}{\cosh^2[c(\alpha y - d\ln 2)]} \tag{3}$$

Table 1. Potential parameters used in the model

Case	D(eV)	$\alpha(\text{Å}^{-1})$	b(eV)	$c(\text{Å}^{-1})$	d	$K(\frac{eV}{\text{Å}^2})$	ρ	$\kappa(\text{Å}^{-1})$
mI: Morse, parameters [19]	AT → 0.05000	4.2	0	\	\	0.025	2	0.35
	GC → 0.07500	6.9	0	\	\			
mII: Morse+barrier [21]	AT → 0.09075	3.0	4	0.74	0.2	0.004	25	0.8
	GC → 0.09900	3.4	6	0.74	0.2			
mIII: barrier 1.5×Ref. [21]	AT → 0.09075	3.0	6	0.74	0.2	0.004	25	0.8
	GC → 0.09900	3.4	9	0.74	0.2			
mIV: barrier 2×Ref. [21]	AT → 0.09075	3.0	8	0.74	0.2	0.004	25	0.8
	GC → 0.09900	3.4	12	0.74	0.2			
mV: Ref. [21] with deep Morse	AT → 0.1255	3.0	4	0.74	0.2	0.004	25	0.8
	GC → 0.1455	3.4	6	0.74	0.2			

Here, $\theta(\cdot)$ is the Heaviside step-function and the original PBD model is reobtained by setting $b = 0$ in the above equation. The heterogeneity of the genetic sequence is taken into account by giving different values to the parameters of the Morse potential for adenine-thymine (AT) or guanine-cytosine (GC) base-pairs. The PBD model does, however, not distinguish between A- and T- nor between G- and C-bases. The additional barrier introduces 3 extra parameters b, c, and d that determine the height, the width, and the position of the added barrier. Together with the original PBD parameters $(D_{\text{AT}}, D_{\text{GC}}, \alpha_{\text{AT}}, \alpha_{\text{GC}}, K, \rho, \kappa)$, the new model is based on 10 parameters in total. Note that this is still significantly less than what is used to fit Poland-Scheraga models. For instance, Ref. [44] reports 31 fitted parameters.

In adapted PBD model, the physically acceptable ranges for values of the width and position of the added barrier are rather narrow. Moreover, since it describes configuration space that is not frequently visited, these two parameters have no dramatic effect on both dynamics and thermodynamics. In our analysis, the c and the d variables

have, therefore, been fixed to the values of Ref. [21] in this work. Moreover, rather than finding a best possible fit to experimental results using a least-square approach, in this work we rather focus on the different characteristics of the different models and investigate whether such a small modification of the PBD model actually allows the accurate description of both dynamics *and* thermodynamics.

The different sets of parameters that we studied are listed in Table 1. Among these are the parameter set of Campa and Giansanti [19, 20] based on the standard PBD model, the original parameters of Ref. [21], and three adaptations of this set. The last three models were studied to detect some trends in the denaturation rate when the shape of the added barrier and the Morse potential is varied and were not a result of intensive parameter fitting. Fig. 1 depicts onsite potentials used in this work.

$$V(y) = D \left(e^{-\alpha y} - 1 \right)^2 + \Theta(y) \frac{by^3}{\cosh^2[c(\alpha y - d \ln 2)]},$$
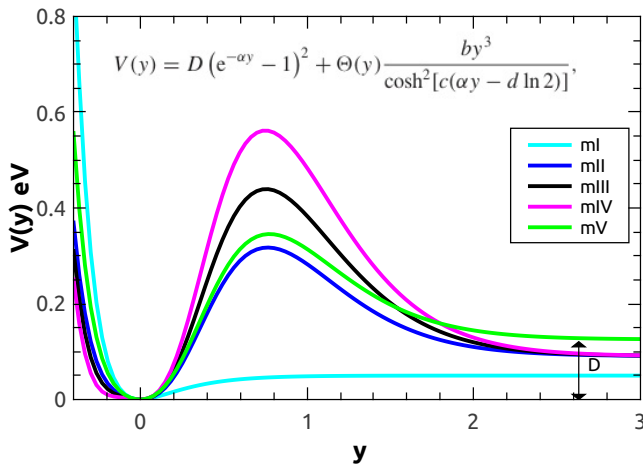
Figure 1. Morse potential used in this work for the models listed in Table 1. The potentials are drawed for an homogenoeus AT sequence.

Like in Refs. [23, 28], the modeling of the DNA hairpin within the PBD framework describes the stem as a double stranded DNA molecule except for the last base-pair before the loop. This terminal base-pair is confined such that its separation y cannot extend a limiting value mimicking the stretch of the loop. In specific, we applied the confining potential of Ref. [23] with a maximum separation for the terminal base-pair of 10 Å. In other words, the loop is not explicitly described but only defines the maximum of the terminal base-pair. It was shown that denaturation curves were not very sensitive to the maximum extension of the last base-pair [23]. In fact, the denaturation curves resemble closely that of double stranded DNA, i.e. without confining the terminal base-pair [23].

3. Rate constant calculation

The reactive flux (RF) approach [45, 46] is a way to compute rate constants of processes that occur at a time-scale that is beyond the accessible time-scale of MD. The method

consists of two independent calculations, a free energy calculation and the calculation of the transmission coefficient. The free energy calculation reveals the free energy profile $F(\lambda)$ as function of a well chosen reaction coordinate λ. In general, the reaction coordinate can be any non-linear function of the particles' coordinates and its value reflects a measure of the progress of the reaction. Hence, in the PBD model, where coordinates are projected on the base-pair separations y_n:

$$\lambda = \lambda(y^N) \text{ with } y^N = \{y_1, y_2, \ldots y_N\} \tag{4}$$

A low value of the reaction coordinate $(\lambda(y^N) < \lambda_A)$ corresponds to reactant state and a high value $(\lambda(y^N) > \lambda_B)$ corresponds to the product state. In between, the transition state surface corresponds to the configurations y^N with a single value of the reaction coordinate $\lambda(y^N) = \lambda_{\text{TST}}$ that is a local maximum of the free energy profile.

The reaction constant within the reactive flux approach can be written as [45]:

$$k_{AB} = \varkappa \sqrt{\frac{k_B T}{2\pi m}} \frac{e^{-\beta F(\lambda^*)}}{\int_{-\infty}^{\lambda^*} e^{-\beta F(\lambda)} \mathrm{d}\lambda} \tag{5}$$

where m is the mass that corresponds the motion of the reaction coordinate and $\beta = 1/(k_B T)$ with k_B the Boltzmann constant and T the temperature. $\varkappa$ is the transmission coefficient which is a number between 0 and 1. The transition state theory (TST) approximation is obtained for $\varkappa = 1$ and $\lambda^* = \lambda_{\text{TST}}$.

TST assumes that each time that the transition state surface is crossed heading towards the product state, it will always reach that state without recrossing the transition state surface another time. The calculation of the transmission coefficient $\varkappa$ (which is explained below) corrects this generally invalid assumption (for the PBD model, the final rate constant is typically a factor 40 lower than than the uncorrected TST expression [6]). The reactive flux method is, therefore, exact and independent of the choice of reaction coordinate or the exact location of the surface λ^*. This is generally set close but not necessarily equal to λ_{TST}.

The free energy $F(\lambda')$ is related to the probability density ϱ by

$$F(\lambda') = -k_B T \ln\left[\varrho(\lambda')/\varrho_0\right] \tag{6}$$

where ϱ_0 is an arbitrary reference density to make the argument of the logarithm dimensionless. The probability density follows from an ensemble average

$$\begin{aligned}\varrho(\lambda') &= \left\langle \delta\left(\lambda(y^N) - \lambda'\right)\right\rangle \\ &= \frac{\int \mathrm{d}y^N \, \delta\left(\lambda(y^N) - \lambda'\right) \exp\left(-\beta\left[\sum_{n=1}^N W(y_n, y_{n-1}) + V(y_n)\right]\right)}{\int \mathrm{d}y^N \, \exp\left(-\beta\left[\sum_{n=1}^N W(y_n, y_{n-1}) + V(y_n)\right]\right)}\end{aligned} \tag{7}$$

where $\delta(\cdot)$ is the Dirac delta function.

To simplify Eq.5 we can introduce the conditional probability density

$$\varrho(\lambda^*|\lambda \leq \lambda^*) = \frac{\left\langle \delta\left(\lambda(y^N) - \lambda^*\right)\right\rangle}{\left\langle \theta\left(\lambda(y^N) - \lambda^*\right)\right\rangle} = \frac{e^{-\beta F(\lambda^*)}}{\int_{-\infty}^{\lambda^*} e^{-\beta F(\lambda)} \mathrm{d}\lambda} \tag{8}$$

$$= \frac{\int \mathrm{d}y^N \, \delta\left(\lambda(y^N) - \lambda^*\right) \exp\left(-\beta\left[\sum_{n=1}^{N} W(y_n, y_{n-1}) + V(y_n)\right]\right)}{\int \mathrm{d}y^N \, \theta\left(\lambda(y^N) - \lambda^*\right) \exp\left(-\beta\left[\sum_{n=1}^{N} W(y_n, y_{n-1}) + V(y_n)\right]\right)}$$

which can be used to substitute the exponential expressions in Eq. 5. The term after the first equals sign is generally the starting point for most common numerical approaches based on molecular dynamics or Monte Carlo. If computational time is not an issue, it can simply be determined by running any of those two simulations and just count how often the reaction coordinate $\lambda(y^N)$ takes a value within a small interval $[\lambda^* : \lambda^* + \mathrm{d}\lambda]$. This number must then be divided by the number of counts that the reaction coordinate was less or equal to λ^* and the width of the interval $\mathrm{d}\lambda$. However, since $[\lambda^* : \lambda^* + \mathrm{d}\lambda]$ interval will be rarely visited or possibly not visited at all during the whole simulation, this small probability must be determined using importance sampling methods like umbrella sampling or thermodynamic integration [45]. Despite being orders of magnitude more efficient than brute-force simulations, these approaches are often still computational intensive as they require the convergence of many different simulations that cover relatively small parts of configuration space.

The alternative is to use the last expression of Eq. 8, but that would require to numerically solve the high-dimensional integrals which is generally unfeasible. The number of grid points, that need to be evaluated, increases exponentially with N and the inverse integration step. However, as shown in Ref. [36], the number of required evaluations can be reduced dramatically using an iterative approach whenever the integrands can be written as product of $N - 1$ two-dimensional functions a_i which only depend on the variables y_i and y_{i-1}, in other words, whenever the integrands are of the form: $a_N(y_N, y_{N-1}) \times a_{N-1}(y_{N-1}, y_{N-2}) \times a_{N-2}(y_{N-2}, y_{N-3}) \ldots \times a_3(y_3, y_2) \times a_2(y_2, y_1)$.

It is henceforth important to chose a reaction coordinate that can both be used to measure the progress of the reaction and will provide the favorable integral expressions. The choice made in Ref. [22] and that we use as well is

$$\lambda(y^N) = \min(y_1, y_2, \ldots, y_N) \tag{9}$$

or the base pair separation of the least open base pair. Using this reaction coordinate Eq. 8 can be written as [22]

$$\varrho(\lambda^*|\lambda \leq \lambda*) = \tag{10}$$

$$\frac{\sum_i \int \mathrm{d}y^N \delta(\lambda(y_i) - \lambda^*) \left(\prod_{j \neq i} \theta(y_j - \lambda^*)\right) \prod_n \exp\left(-\beta\left[W(y_n, y_{n-1}) + V(y_n)\right]\right)}{\int \mathrm{d}y^N (1 - \prod_k \theta(y_k - \lambda^*)) \prod_n \exp\left(-\beta\left[W(y_n, y_{n-1}) + V(y_n)\right]\right)}$$

Both the nominator and the denominator of the above expression can be written as a sum of integrals that all have the favorable product form and can, hence, be solved using Simpson's rule and the efficient iterative integration method.

The second calculation is to determine the transmission coefficient $\varkappa$, which requires the release of many molecular dynamics trajectories with starting points picked from a equilibrium distribution at the surface $\lambda(y^N) = \lambda^*$. The description above is generic for any reactive flux approach, but there are still some variations based on equivalent mathematical formulations of $\varkappa$ [47].

It is most convenient to introduce the *unnormalized* transmission coefficient R

$$R = \varkappa \sqrt{\frac{k_B T}{2\pi m}} \tag{11}$$

So that Eq. 5 is a simple product

$$k_{AB} = R\varrho(\lambda^* | \lambda \leq \lambda^*) \tag{12}$$

Now, as discussed in Ref. [47] the different expressions for R can all be written as

$$R = \frac{\left\langle \dot{\lambda}\chi(y^N, p^N)\delta\left(\lambda(y^N) - \lambda^*\right) \right\rangle}{\left\langle \delta\left(\lambda(y^N) - \lambda^*\right) \right\rangle} = \left\langle \dot{\lambda}\chi(y^N, p^N) \right\rangle_{\lambda^*} \tag{13}$$

where the subscript λ^* at the bracket implies that the ensemble average has to be taken over configurations with $\lambda(y^N) = \lambda^*$. $\dot{\lambda}$ is the time-derivative of λ or the velocity along the reaction coordinate of the phase point (y^N, p^N) that is at the surface λ^*. The function $\chi(y^N, p^N)$ has usually two possible outcomes, 0 or 1, and its value is determined by a short MD trajectory starting from the phase point (y^N, p^N). Like in Ref. [27], we used the *effective positive flux* formalism which requires to go both forward and backward in time. The later is achieved by simply reverting the momenta. The expression for χ in this formalism is

$$\chi(y^N, p^N) = \begin{cases} 1 & \text{if the MD trajectory starting from } (y^N, p^N) \text{ hits } \lambda_B \\ & \text{before } \lambda_A \text{ and if the MD trajectory starting from} \\ & (y^N, -p^N) \text{ hits } \lambda_A \text{ before } \lambda^*. \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Note that whenever $\dot{\lambda} < 0$ the backward trajectory cannot cross λ_A before crossing λ^*. Hence, χ can be assigned 0 without actually running an MD trajectory in that case. If $\dot{\lambda} > 0$, the backward trajectory is performed first and then the forward trajectory only if λ_A was reached before λ^*. Hence, if the backward movement crosses λ^*, χ is directly assigned 0 saving the need of many additional MD steps.

The effective positive flux expression does not only reduce the number of MD steps per χ-evaluation, it also reduces the number of evaluations needed. Since $\dot{\lambda}\chi$ is always greater or equal to zero for any phase point, the effective positive flux expression also converges faster than other expressions where $\dot{\lambda}\chi$ can be both positive and negative [47]. Note that if the TST assumption is valid Eq. 14 reduces to $\chi = \theta(\dot{\lambda})$ and Eq. 13 can be solved analytically resulting in $R = \sqrt{k_B T/2\pi m}$ as expected from Eq. 11.

Another algorithmic trick [22] that optimally utilizes specific characteristics of PBD-class models allows the effective generation of uncorrelated starting points (y^N, p^N) at the λ^* surface. This trick is based on the fact that the statistical ensemble with Hamiltonian H of Eq. 1 that is constrained to the surface $\lambda(y^N) = \lambda^*$ is equivalent to that of a freely moving chain on a translational invariant Hamiltonian H' which relates to the true Hamiltonian H via

$$H'(y^N, p^N) = H(y^N - \lambda(y^N) + \lambda^*, p^N) \tag{15}$$

Henceforth, the generation of surface points can be done by running an unconstrained MD simulation using Hamiltonian H' and store the generated phase point every Mth

time step, where M is chosen large enough to reduce correlations. Then, each stored point is first shifted by adding a single constant to all y_i values such that its minimum as at λ^*, and then MD trajectories are generated with the correct Hamiltonian H in order to compute χ of that point.

We applied this method using a numerical integration step of $dy = .05$ Å for the numerical free energy calculation. Integration boundaries were set according to Ref. [6] using a tolerance of $\epsilon = 10^{-40}$, which implies that contributions with Boltzmann weight $e^{-\beta H}$ lower than this value are neglected. The phase points on the surface were generated using Hamiltonian H' and saving every $M = 1000$th step. For each rate calculation, we generated 10^6 surface points which were used to calculate χ based on $\lambda_A = 0$, $\lambda^* = 0.75$, and $\lambda_B = 2.5$ Å. Both for the generation of points using H' and the evaluation of χ using H the dynamics were influenced by a heat-bath with the requested temperature via a Langevin thermosthat using a time step of 1 fs and a base-pair mass m of 300 amu. The Langevin friction was set to 50 ps^{-1} which is considered to give accurate dynamics for biomolecules in water [48]. For friction coefficients larger than 10 ps^{-1}, the transmission coefficient and the rate constant of the standard PBD model become inversely proportional to the friction [6]. This agrees well with Kramer's theory in the high-friction regime [49]. Hence, for a friction of 50 ps^{-1} it seems well justified to rely on overdamped Langevin dynamics (Brownian dynamics) in which the inertia terms are ignored. Brownian dynamics could provide a more efficient way to evolve the time evolution of the system, but it is not compatible with the standard approach for calculating transmission coefficients [50]. It is therefore that we use underdamped Langevin equations of motion, even if the friction is relatively high. For an extended discussion about the friction parameter ranges used in PBD models we refer to Ref. [51].

4. Results and discussion

Fig. 2 shows the results obtained from the methodology of Refs. [6,22,23] and the model of Eqs. 1- 2 of the calculated hairpin denaturation rate constants for a set of different sequences for which experimental values are available.

Table 2 show the sequences and temperatures of the different data points in Fig. 2 together with their experimental reference. Table 1 shows the parameter sets used in Eq. 2. Each parameter set defines a case model, notated as mI-mV. In Fig. 2, only the models mI and mII have been shown corresponding to, respectively, the original PBD model [18] with parameters from Campa and Giansanti [19] and the model with the additional barrier of Ref. [21]. Note that the model parameters of Ref. [21] were fitted to experimental data on base-pair life-times and not to denaturation rates while Ref. [19] was fitted to thermodynamic data alone.

Clearly, the adapted PBD model (mII) shows considerably better agreement with the experimental results than the original barrierless PBD model(mI). Whereas the latter is three till five orders of magnitude too high, the adapted model (mII) is either on top of the results for the short sequences (5 base-pair stem) or maximally two orders
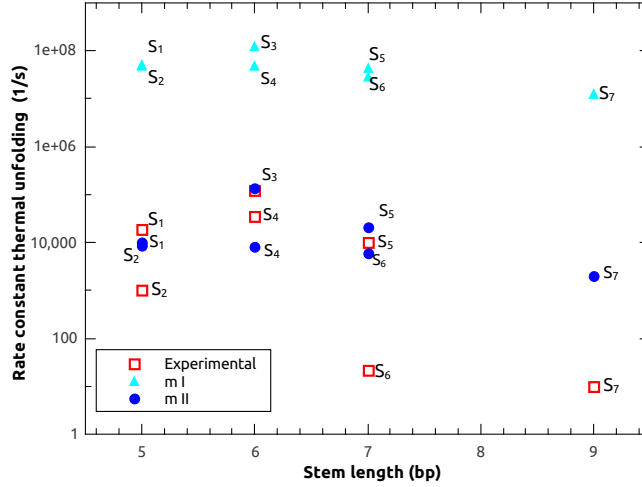
Figure 2. Thermal opening rates of hairpins from Refs. [27, 29, 30, 32–34]. Shown are the results of the original barrierless PBD model (mI) and PBD model with barrier (mII) compared with the experimental results for sequences listed in Table 2. Statistical errors of the simulations are within symbol size.

Table 2. a) Sequences and temperatures of thermal unfolding hairpins. b) Sequences of hairpins used in optical tweezer experiments.

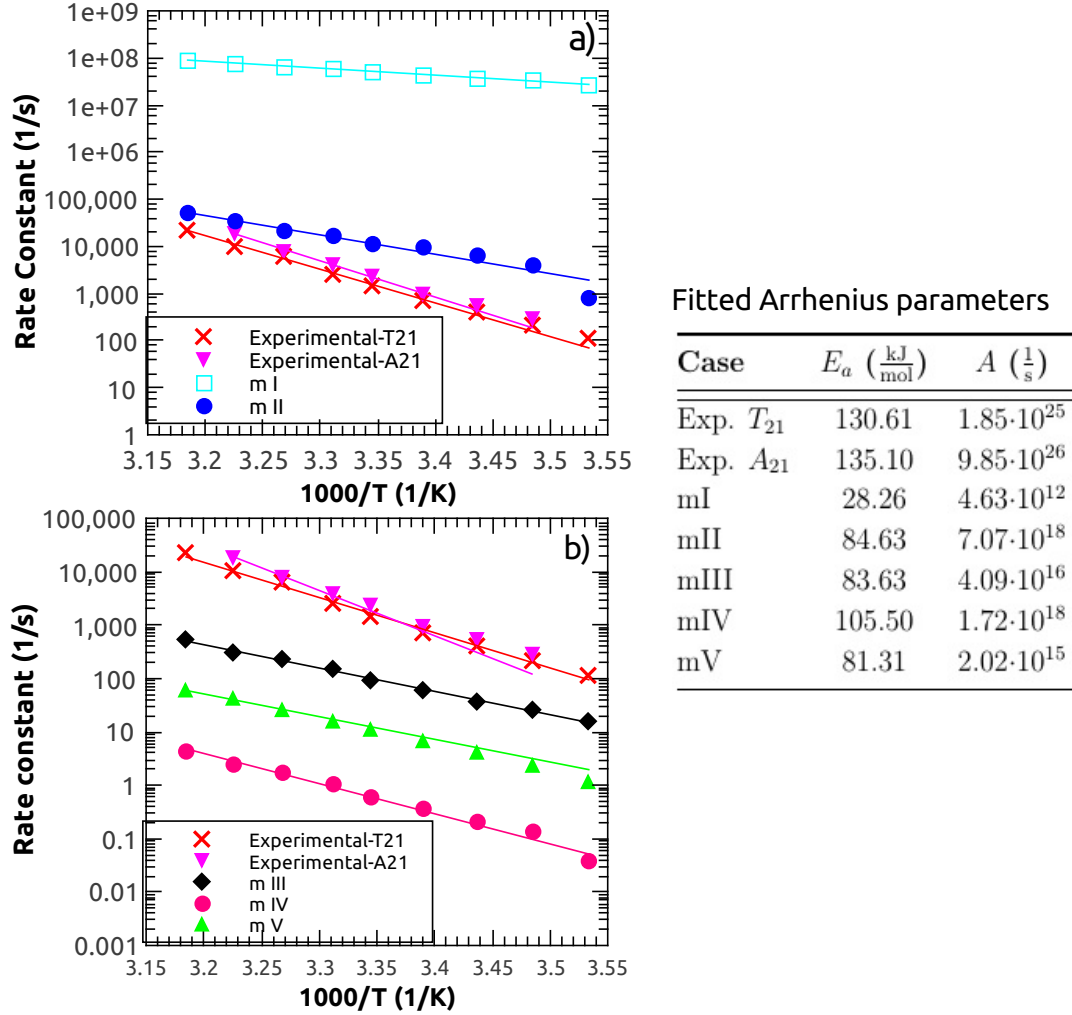| **a) Thermal denaturation exp. Ref. [27, 29, 30, 32–34]** | | | | |
Label	**Stem**	**Loop**	**Temp.(°C)**	**Ref.**
S_1	AAGGG	T_{21}	25	[29]
S_2	GGGAA	$T_{n=4};A_{30}$	23	[27,33]
S_3	AAGGAA	T_4	54	[30]
S_4	GGAAAA	T_4	23	[31]
S_5	GGGAAAA	T_8	36	[32]
S_6	GAGAAGA	A_{40}	25	[34]
S_7	GAGAAGAGA	A_{40}	25	[34]
b) Optical tweezer exp. Ref. [35]				
Length	**Stem**		**Temp.= 23 °C**	
6	GAGGAA			
8	GAGAGGAA			
10	GAGAGAGGAA			
15	GAGAGGAGGAAGGAA			
20	GAGAGAAGGAGAGGAAGGAA			
25	GAGAGAAGGAAGAGGGAGGAAGGAA			
30	GAGAGAAGGAAGAGAAGAGGGAGGAAGGAA			

of magnitude off for the longer ones (9 base-pair). It is important to realize that also the experimental data show huge variations for sequences having the same stem length and

same GC content. This can be due to the differences in loop length or sequence, or due to variations in the temperatures under which the experiments were performed. However, it is also not unthinkable that the different experimental measurement techniques perturb the dynamics and, therefore, possibly won't provide the same rates. We come back to this when we compare our results with that of another experimental study based on optical tweezers [35].

In Fig. 3 we examined the denaturation rate as function of temperature. Experimental results were taken from Ref. [27] by Bonnet *et* al. who investigated two hairpins with identical five base-pair stem GGGAA but with loops either consisting of 21 A-bases or 21 T-bases. Note that these two systems have identical descriptions in our models, not only because it doesn't distinguish between A and T bases, but also because the loop itself is not explicitly described as explained above. In Fig. 3-a) we use models mI and mII of table 2 which do not involve any parameter fitting. The results indicate that the original PBD model deviates from the experimental results by five and three orders of magnitude at low (ca. 280 K) and high temperatures (ca. 315 K), respectively. The model with additional barrier shows for the whole range of temperatures results that are within one order of magnitude of the experimental values. From the slope and the intersection of the straight fitted lines, we obtained the Arrhenius parameters, the activation energy E_a and the pre-exponential factor A, which are shown in Fig. 3-c). It shows that the mismatch between theory and experiment is largely due the pre-exponential factor.

Overall, this shows that model proposed in Ref. [21] considerably improves the dynamics of DNA and, especially for short DNA sequences, reproduces the experimental denaturation rate constants reasonably well. Fig. 3-b) shows the same experimental results again together with the models mIII-mIV. Model mIII is nearly as good as mII with the difference that mIII gives rate constant that are lower than the experimental results, while the mII results higher. On the logarithmic scale, the curves of the rate constants as function of the inverse temperature have nearly the same slopes.

On the other hand, if we compare the rate versus stem length we see a different behavior. Fig. 4 shows for different stem length the denaturation rate of seven experimental DNA hairpins. The stem sequences are given in Table 2-b) and contain an 50% GC base-pairs and a 4T loop sequence at a temperature of $23\,^\circ$C. The experimental denaturation rates are taken from the data of unloaded (in absence of external force) optical tweezer experiments [35] which show a steep exponential decay as function of chain length. The PBD models mI and mII produce a much weaker decay as function of chain length which has as a consequence that the deviation between the model and the experimental results increase beyond 15 orders of magnitude for the longest chain in this data set (30 base-pairs). The model with barrier (Ref. [21]) is around 5 orders of magnitude lower than the original PBD model for all sequences, i.e. closer to the experimental results, but it fails to reproduce the same slope in the $\ln k$ versus chain length plot. To investigate whether the shape of the barrier could be adjusted to reproduce the same experimental trend we looked at the models mIII and mIV (See

Fitted Arrhenius parameters

Case	$E_a \left(\frac{\text{kJ}}{\text{mol}}\right)$	$A \left(\frac{1}{\text{s}}\right)$
Exp. T_{21}	130.61	$1.85 \cdot 10^{25}$
Exp. A_{21}	135.10	$9.85 \cdot 10^{26}$
mI	28.26	$4.63 \cdot 10^{12}$
mII	84.63	$7.07 \cdot 10^{18}$
mIII	83.63	$4.09 \cdot 10^{16}$
mIV	105.50	$1.72 \cdot 10^{18}$
mV	81.31	$2.02 \cdot 10^{15}$

Figure 3. Opening constant of hairpins with stem GGGAA and loop of T21 and A21 as a function of the inverse temperature (from Ref. [27]) and comparison with our model with different sets of parameters. a) Experimental data compared with theoretical results from the original PBD model (mI) with parameters from Refs. [19, 20] and the adapted PBD model (mII) with parameters from Ref. [21]. b) Comparison of the same experimental data with the theoretical results of the adapted PBD model with parameter variations (models mIII-mV of table 1)

Table 1) in which we increase the barrier height by a factor 1.5 and 2, respectively.

Clearly, the barrier increase gives an overall shift in the logarithmic plot but hardly changes the slope. The reason for this is that this increase has two opposing effects: Not only will it slow down the rate of individual base-pairs to open, it also makes it more difficult for this base-pair to reclose. The latter effect increases the denaturation rate of the whole DNA sequence and is the reason for the decay being independent of the barrier height (parameter b in Eq. 2). In order to get qualitatively correct behavior of the denaturation rate versus chain length, there is no other option than to change the relative probabilities between the open and closed states for the individual base-pairs. Equivalently, the barrier of opening should be increased without increasing the
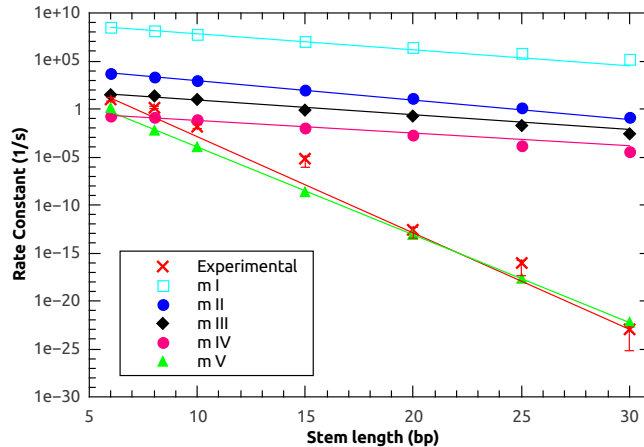
Figure 4. Rate constant of denaturation measured by unloaded optical tweezer experiments [35] and calculated by thermal denaturation simulation using the models mI-mV of Table 1.

closure barrier with the same amount. To do this, we changed the depth of the Morse potential (or equivalently, as the minimum is set at zero, we increased the height of the plateau) which lead to model mV which can quantitatively reproduce the optical tweezer experiments at all lengths.

Changing the height of the Morse potential comes with a prize as can be seen in Fig. 5 where we show the denaturation curves of long (400 base-pairs) homopolymers which shows the typical phase transition behavior [18]. Experimentally this transition occurs around 70 °C for pure GC sequences [52] (with variations depending on the ion concentration), but model mV gives melting transitions for AT and GC chains that are beyond the boiling temperature of water. Hence, this shows that it is impossible to get both the thermodynamics and dynamics correct at the same time using a single unified mesoscopic model. Improvements might lie in the replacement of the flat horizontal Morse plateau with a function having negative slope to describe entropic effects yet unconsidered such as rotational degrees of freedom and the possibility for open regions to curl up. Ideally, the development of such a model should be derived from full atomistic simulations but the computational expense for simulating long DNA sequences in explicit water is yet too large. The alternative is to use a more complex model and fit it to experimental data. This should be done on both thermodynamic and dynamical data. Unfortunately, the experimental uncertainties regarding denaturation rates are still very large which makes it difficult to decide which characteristics the theoretical models should have. The deviation in experimental results is obvious from Fig. 2. The discrepancy is even larger if we compare those results with the results of the optical tweezer experiments with zero load [35] of Fig. 4 for the short sequences. This comparison shows that the optical tweezer experiments give results 2 till 3 orders of magnitude lower compared to the thermal denaturation experiments for short chain lengths. For stem lengths of 6 base-pairs, the optical tweezer experiment gives results

close to that of models mIII, mIV, mV (Fig. 4) However, for longer stem lengths, only the model with the deep Morse potential can reproduce the optical tweezer experiments.
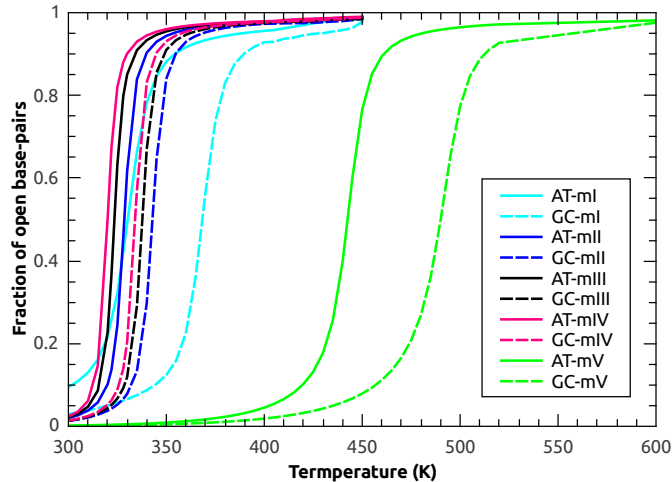
Figure 5. Fraction of open base-pairs as function of temperature for long (400 base-pairs) homopolymers calculated using the approach of Ref. [5] with models mI-mV of Table 1

.

5. Conclusion

We showed that a recently adaptation [21] of the PBD model with an additional barrier to the Morse potential improves the agreement of the denaturation rate for DNA hairpins in comparison with experimental results. For short sequences, this agreement is relatively good, but the magnitude of the exponential decay of the rate as function of chain length does not correspond to those measured by optical tweezer experiments [35]. As a results, the overestimation of denaturation rates of the model in comparison to experiments increases with increasing chain length. As we show, it is possible to change the model parameters such that they reproduce the right decay, but this increases the melting temperature of the long homopolymers to unrealistic values. Therefore, a more complex model should ideally be fitted to both thermodynamic and dynamical data, especially since the latter are much more sensitive to model parameters. The data could be denaturation rates, but also other dynamical properties could be used for fitting like thermal conductance [9]. Possibly a slightly more complex helicoidal model [53] that more explicitly includes the effects of bending, torsional energies, and unwinding, would be more successful in describing both dynamics and thermodynamics at the same time. Another possibility that could resolve the mismatch between thermodynamics and dynamics is to let the model parameters be temperature dependent. However, fitting such a model, while maintaining some kind of physical justification, is far from trivial.

An essential condition for establishing better models is an increase in both quantity and quality of experimental data on dynamics. Rate constants are more difficult to measure in comparison with thermodynamic properties and, correspondingly, error-bars are typically a few order orders of magnitude. For a long time, rate constants from computer simulations were equally inaccurate, but due to the increasing computer power and algorithmic innovations this has changed. Especially for first-neighbor type of models, rate constants can be obtained with very low statistical errors using the method described in this article. Therefore, a similar improvement in the experimental accuracy would be very welcome and allow for a more reliable parameter fitting of theoretical models.

More disturbing from a theoretician's point of view is that experimental rate constants based on different experimental techniques do not always agree, even when considering the large error-bars. This field would therefore benefit from more systematic or automated experimental approaches. Defining a few standardized reference conditions related to ion concentrations and temperature would facilitate the comparison of results by different experimental groups. High-throughput screening could eliminate human factors in e.g. sample preparation and reduced error-bars due to an increase of repetitive identical measurements. In addition to this, it would be good to find estimates about the effect of the experimental probe on the results, and on the validity of some theoretical assumptions, like the infinite dilution limit which neglects interactions between different DNA molecules. Finally, the design of theoretical models would be highly facilitated by an increase of experimental data on synthetically created 'simple' DNA sequences, like homogeneous A- or G-sequences of different length, alternating (AGAGAG...) sequences, and the ones reported in Refs. [6,23]. We therefore hope that our results will inspire both experimentalists and theoreticians to study dynamical transitions of DNA.

Acknowledgments

6. References

[1] Cuesta-Lopez S, Peyrard M and Graham D 2005 *Eur. Phys. J. E* **16** 235–246
[2] Watson J D, Crick F H *et al.* 1953 *Nature* **171** 737–738
[3] Peyrard M and Bishop A R 1989 *Phys. Rev. Lett.* **62** 2755–2758

[4] Alexandrov B, Voulgarakis N K, Rasmussen K Ø, Usheva A and Bishop A R 2009 *J. Phys.: Condens. Matter.* **21** 034107

[5] van Erp T S, Cuesta-Lopez S and Peyrard M 2006 *Eur. Phys. J. E* **20** 421–434

[6] van Erp T S and Peyrard M 2012 *EPL* **98** 48004

[7] Markham N R and Zuker M 2005 *Nucleic Acids Res.* **33** W577–W581

[8] Takinoue M and Suyama A 2006 *Small* **2** 1244–1247

[9] Chien C C, Velizhanin K A, Dubi Y and Zwolak M 2013 *Nanotechnology* **24** 095704

[10] Ising E 1925 *Zeitschrift für Physik A Hadrons and Nuclei* **31** 253–258

[11] SantaLucia J 1998 *Proc. Natl. Acad. Sci. U. S. A.* **95** 1460–1465

[12] Poland D and Scheraga H A 1966 *The Journal of chemical physics* **45** 1456–1463

[13] Benham C J 1979 *Biopolymers* **18** 609–623

[14] Fye R M and Benham C J 1999 *Physical Review E* **59** 3408

[15] Ivanov V, Zeng Y and Zocchi G 2004 *Physical Review E* **70** 051907

[16] Kittel C 1969 *American Journal of Physics* **37** 917–920

[17] Dasanna A K, Destainville N, Palmeri J and Manghi M 2013 *Physical Review E* **87** 052703

[18] Dauxois T, Peyrard M and Bishop A R 1993 *Phys. Rev. E* **47** R44

[19] Campa A and Giansanti A 1998 *Phys. Rev. E* **58** 3585

[20] Campa A and Giansanti A 1999 *J. Biol. Phys.* **24** 141–155

[21] Peyrard M, Cuesta-Lopez S and Angelov D 2009 *J. Phys.: Condens. Matter* **21** 034103

[22] van Erp T S 2007 *Phys. Rev. Lett.* **98** 268301

[23] Dahlen O and van Erp T S 2015 *J. Chem. Phys* **142**

[24] Zazopoulos E, Lalli E, Stocco D M and Sassone-Corsi P 1997 *Nature* **390** 311–315

[25] Trinh T Q and Sinden R 1993 *Genetics* **134** 409

[26] Proctor D J, Ma H, Kierzek E, Kierzek R, Gruebele M and Bevilacqua P C 2004 *Biochemistry* **43** 14004–14014

[27] Bonnet G, Krichevsky O and Libchaber A 1998 *Proc. Natl. Acad. Sci. USA* **95** 8602–8606

[28] Hanne J, Zocchi G, Voulgarakis N K, Bishop A R and Rasmussen K Ø 2007 *Phys. Review E* **76** 011909

[29] Jung J and Van Orden A 2005 *J. Phys. Chem. B* **109**

[30] Hilbers C W, Haasnoot C A G, De Bruin S H, Joordens J J M, Van Der Marel G A and Van Boom J H 1985 *Biochimie* **67** 685–695

[31] Ansari A, Kuznetsov S V and Shen Y 2001 *Proc. Natl. Acad. Sci. USA* **98**

[32] Ansari A and Kuznetsov S V 2005 *J. Phys. Chem. B* **109** 12982–12989

[33] Wallace M I, Ying L, Balasubramanian S and Klenerman D 2001 *Proc. Natl. Acad. Sci. USA* **98**

[34] Grunwell J R, Glass J L, Lacoste T D, Deniz A A, Chemla D S and Schultz P G 2001 *J. Am. Chem. Soc.* **123** 4295–4303

[35] Woodside M T, Behnke-Parks W M, Larizadeh K, Travers K, Herschlag D and Block S M 2006 *Proc. Natl. Acad. Sci. USA* **103** 6190–6195

[36] van Erp T S, Cuesta-Lopez S, Hagmann J G and Peyrard M 2005 *Phys. Rev. Lett.* **95** 218104

[37] Theodorakopoulos N and Peyrard M 2012 *Phys. Rev. Lett.* **108** 078104

[38] Leroy J L, Kochoyan M, Huynhdinh T and Gueron M 1988 *J. Mol. Biol.* **200** 223–238

[39] Weber G 2006 *EPL (Europhysics Letters)* **73** 806

[40] Peyrard M and James G 2012 *Nonlinear Theory and Its Applications, IEICE* **3** 27–51

[41] Drukker K, Wu G S and Schatz G C 2001 *J. Chem. Phys.* **114** 579–590

[42] Giudice E, Varnai P and Lavery R 2003 *Nucleic Acids Res.* **31** 1434–1443

[43] Valle-Orero J, Wildes A R, Theodorakopoulos N, Cuesta-López S, Garden J L, Danilkin S and Peyrard M 2014 *New J. Phys.* **6** 113017

[44] Jost D and Everaers R 2009 *Biophys. J.* **96** 1056–1067

[45] Frenkel D and Smit B 2002 *Understanding molecular simulation, 2nd ed.* (San Diego, CA: Academic Press)

[46] Chandler D 1987 *Introduction to Modern Statistical Mechanics* (New York: Oxford University

Press)

[47] Van Erp T S 2012 *Advances in Chemical Physics* **151** 27

[48] Pastor R W, Brooks B R and Szabo A 1988 *Mol. Phys.* **65** 1409–1419

[49] Hanggi P, Talkner P and Borkovec M 1990 *Rev. Mod. Phys.* **62** 251–341

[50] Menzl G, Singraber A and Dellago C 2016 *Faraday Discuss.* **195** 345–364

[51] Manghi M and Destainville N 2016 *Phys. Rep.-Rev. Sec. Phys. Lett.* **631** 1–41

[52] Inman R and Baldwin R 1964 *Journal of molecular biology* **8** 452–469

[53] Barbi M, Cocco S and Peyrard M 1999 *Phys. Lett. A* **253** 358–369