PHYSICAL REVIEW E **74**, 050901(R) (2006)

Bubble statistics and dynamics in double-stranded DNA

B. S. Alexandrov,[1,2,3] L. T. Wille,[3] K. Ø. Rasmussen,[1] A. R. Bishop,[1] and K. B. Blagoev[1]
[1]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2]*Nuclear Nonproliferation Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*Physics Department, Florida Atlantic University, Boca Raton, Florida 33431, USA*
(Received 23 January 2006; published 1 November 2006)

The dynamical properties of double-stranded DNA are studied in the framework of the Peyrard-Bishop-Dauxois model using Langevin dynamics. Our simulations are analyzed in terms of two distribution functions describing localized separations ("bubbles") of the double strand. The result that the bubble distributions are more sharply peaked at the active sites than thermodynamically obtained distributions is ascribed to the fact that the bubble lifetimes affect the distributions. Certain base-pair sequences are found to promote long-lived bubbles, and we argue that this is a result of length scale competition between the nonlinearity and disorder present in the system.

DOI: 10.1103/PhysRevE.74.050901 PACS number(s): 87.14.Gg, 05.45.Pq

The role of dynamics in biological function is becoming increasingly clear [1–3]. Whereas protein action and binding have traditionally been discussed in terms of static structures, it is now evident that many functionalities are consequences of dynamics. Because of its biological importance and structural clarity, DNA constitutes an appropriate system in which to begin to understand how structure and thermal motion can work together to determine function. In particular, the identification of biological processes that are regulated by the dynamical properties of DNA is fundamentally important for understanding its interaction with other molecules. Key mechanisms may be entropically driven thermal fluctuations, which cause local dynamical changes in the interstrand separation ("bubbles") in double-stranded DNA molecules. Recent theoretical and experimental studies [4] suggest that the base-pair sequence (structure) determines specific regions in the double strand that are more prone to such thermally induced strand separation. Most importantly, these studies have demonstrated a strong correlation between the specific location of large, coherent openings in DNA and transcription-promoting regions of the DNA sequence for several well-characterized viral sequences. It has also be found [5] that UV-induced dimers of thymine bases dramatically altered the DNA dynamics. This suggests an important role of large fluctuations in the dimer recognition pathway. In both cases the theoretical characterization has been provided by the Peyrard-Bishop-Dauxois (PBD) model [6,7]. However, recently it was argued [8,9] that thermodynamic characterization of the thermal fluctuations may differ from a dynamical characterization, which points to the need for a thorough understanding of the dynamical effects in this highly nonlinear and cooperative material.

Here, we use finite-temperature Langevin simulations to probe the impact of sequence heterogeneity on bubble dynamics in six different sequences, all composed of 69 base pairs: (i) two homogeneous sequences composed purely of thymine-adenine (T-A) and guanine-cytosine (G-C) base pairs, respectively, and (ii) two specific heterogeneous sequence: Adeno Associate Viral (AAV) P5 promoter and a mutated (AAV) P5 promoter (see Ref. [4] for details). Finally, (iii) we investigate two periodic sequences, each containing 35 (T-A) base pairs and 34 (G-C) base pairs that have different periodicities—G_1A_1 with a period of 2 base pairs and G_5A_5 with a period of 10 base pairs. We compare our results to thermodynamic results for the interstrand DNA opening obtained with the same model and observe several important differences.

It should be emphasized that the Langevin dynamics of the PBD model's principal degrees of freedom for DNA base pairs is necessarily a phenomenological representation of DNA's full complexity: microscopic fine scales of, e.g., water motions are not explicitly modeled. Similarly are large-scale effects and the resulting dynamical time scales, caused by three-dimensional conformations and torsions of DNA molecule, intrinsically absent because of the one-dimensional character of the model. In this sense the PBD model yields, as far as concerns dynamics, a more qualitative than quantitative description. This most clearly manifests itself in the discrepancy between our calculated bubble opening times and the experimentally inferred opening times [10]. Some of this discrepancy can further be ascribed to the fact that the correct damping constant is largely unknown and we are forced to use a value for this parameter that is known to leave the thermodynamical properties unaffected. Nevertheless, the model does give a qualitative idea of the dynamical issues involved in the bubble openings, which leads us to argue that the dynamics may play a more important role in the prediction of biologically important sites than the thermodynamic quantities do.

The thermal dynamics of the nth base pair is obtained through

$$m\ddot{y}_n = -V'(y_n) - W'(y_{n+1},y_n) - W'(y_n,y_{n-1}) - m\gamma\dot{y}_n + \xi_n(t), \tag{1}$$

where $V(y)=D_n(e^{-a_n y}-1)^2$ is an on-site Morse potential modeling the hydrogen bonding of complementary bases and representing the exact sequence (see Ref. [4]) and $W(x,y)=k/2(1+\rho e^{-\beta(x+y)})(x-y)^2$ represents the nonlinear stacking interactions. Here a prime denotes differentiation with respect to y_n, γ is the friction constant, and the random force $\xi_n(t)$ is Gaussian-distributed white noise. With the parameter values used (see Refs. [11,12]), the success of the

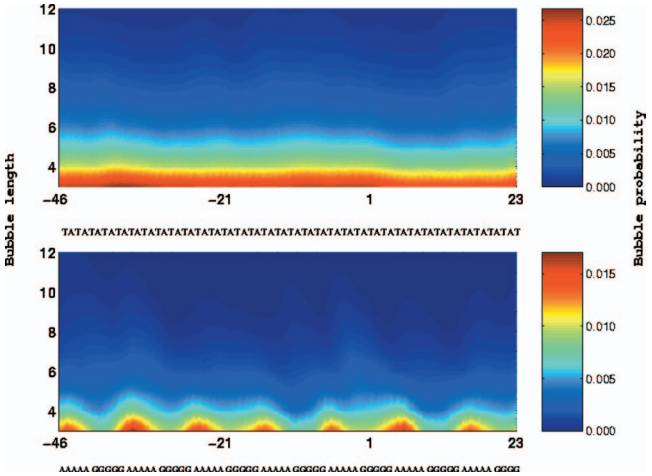

FIG. 1. (Color) Bubble probability $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ for the 69-base-pair homogeneous A-T sequence (upper) and for the periodic G_5A_5 sequence (lower).

model in describing the base-pair openings of double-stranded DNA has been demonstrated by direct comparison with various experiments on the melting transition [12], S1 nuclease digestion [4], predenaturation bubbles [13], and forced unzipping [14].

Here we simulate the dynamics of double-stranded DNA at $T=300$ K by numerically integrating the system of stochastic equations (1), applying periodic boundary conditions. In the presence of the thermal bath, modeled by the random forces, the creation of a bubble is a stochastic process [15] most appropriately described in terms of a probability. We define the probability for bubble existence as

$$P_n(l,tr) = \left\langle \frac{1}{t_s} \sum_{q_n^k=1}^{q_n^{k_{max}}(l,tr)} \Delta t[q_n^k(l,tr)] \right\rangle_M , \qquad (2)$$

where $\langle \cdots \rangle_M$ denotes averaging over M ($\sim$1000) simulations. The integer index $q_n^k(l,tr)$ enumerates the bubbles defined as a double-strand separation of amplitude $tr \geqslant 0.5$ Å, spanning $l \geqslant 3$ consecutive base pairs beginning at the nth base pair in the kth simulation. In practice we bin $P_n(l,tr)$ using bin sizes $l=1$ and $tr=0.5$ Å. The quantity $\Delta t[q_n^k(l,tr)]$ is the existence time of the $q_n^k(l,tr)$th bubble. $t_s \sim 1-2$ nsec is the duration of a single simulation [bubble lifetimes are in the picosecond range and are much smaller see (Fig. 4)]. Probabilities for bubble existence at a given site, for all lengths and amplitudes defined in this way, are obviously normalized since all possible openings at every step of the simulation are counted. The plot of $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ as obtained from Eq. (2) given in Fig. 1 demonstrates two clear results.

(i) The probabilities for bubble formation in a homogeneous T-A sequence (or for a G-C sequence, not shown) do not depend on the base-pair index because of the translation invariance.

(ii) For the periodic sequence G_5A_5, the probabilities are periodic with the period of the sequence. We can clearly identify the *sources* of the bubbles to be situated in the

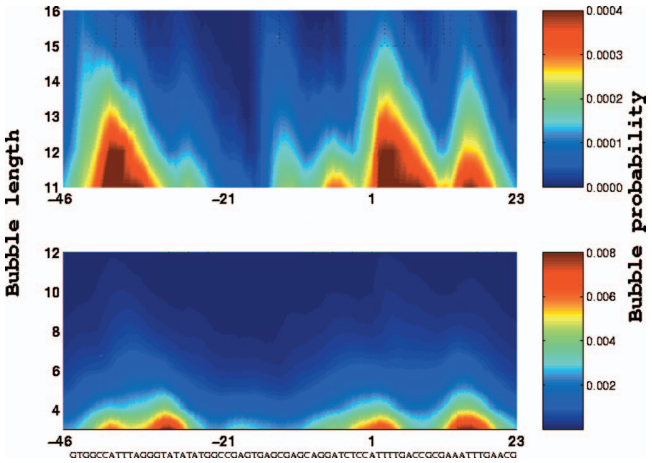

FIG. 2. (Color) Bubble probability $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ for the 69-base-pair AAV P5 promoter (see Ref. [4]), with $t_s=2$ nsec and $M=800$ simulations.

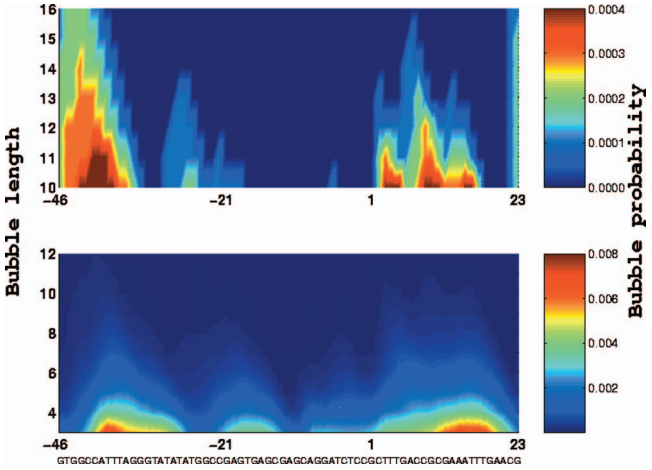

FIG. 3. (Color) Bubble probability $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ for a mutated AVV P5 sequence (see text).

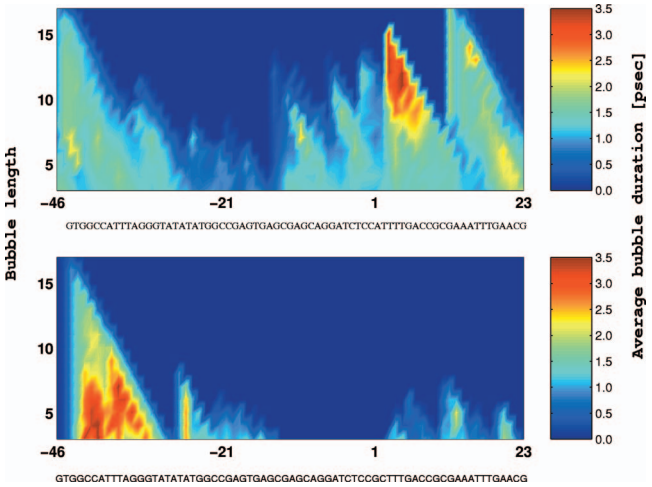

FIG. 4. (Color) Average bubble duration time $\Sigma_{tr=4.5\ \text{Å}}^{\infty} ABD(n,l,tr)$ for AAV P5 (upper) and for mutated AVV P5 (lower) sequences.

AT-rich half-period of the sequence. In contrast, we observed for a G_1A_1 sequence (not shown) the probabilities to be almost spatially uniform because of the short periodicity—only a single G-C base pair between the two A-T base pairs. The important observation from these results is that not only the length of the hot spots (the T-A areas) but also the length of the "barriers" (the G-C intervals) play crucial roles for the probability of the bubble existence by restricting, through impedance mismatch, the energy flow from the (A-T)-rich regions. Clearly bubbles of all length have thermodynamic weight, increasing with temperature and decreasing with bubble size. However, the base-pair inhomogeneity preferentially selects long-lived bubbles of specific sizes at specific locations. This is a result of length-scale competitions inherent in nonlinear, disordered systems [16].

In Fig. 2 we show $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ for the 69 base-pair adeno associate viral (AAV) P5 promoter (see Ref. [4]). Two regions located around base +1 and −30 are prominent in terms of largest probabilities for bubbles existence. It is noteworthy that the probability becomes more localized around the identified sites as increasingly longer (in terms of consecutive sites) bubbles are considered. This result is in agreement with findings in similar simulations investigating the frequency of the base-pair opening [4]. There is also good agreement with the thermodynamic results [8,9] derived from the PBD model; however, the active regions are much more sharply identified here than in the case of the thermodynamic treatment.

Our results confirm that the thermal bubble localized at the transcription promoter site can aid the RNA polymerase and the associated proteins in the formation of the transcription bubble

It has been argued [8] that the results obtained in Ref. [4] are flawed by insufficient statistics in the simulations. We therefore present in Fig. 2 the result of $M=800$ simulations of $t_s=2$ nsec duration. Similar results were obtained from simulations of $t_s=1$ nsec duration (not shown), suggesting that the statistics is indeed sufficient even in a 1-nsec simulation. It is important to note that even for the 2-nsec simulations we never observed complete melting of all the 69 base pairs, indicating that we are exclusively sampling the premelting regime.

In Fig. 3 we similarly show $\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ for a mutated AVV P5. The mutation, which has severe consequences for the promoter's ability to induce transcription (see Ref. [4]), consists of changing base pairs +1 and +2 to G-C pairs. From Fig. 3 we observe that this mutation indeed severely inhibits the formation of large bubbles around the +1 base pair. Comparing Figs. 2 and 3, we see that changing just two base pairs is a sufficient increase of the G-C "barrier" to restrict the flow of thermal energy to be exclusively downstream in the sequence. This is the mechanism by which the mutation can induce a rather long-range effect, such as a change in the probability around base pair −30, although this effect may be specific to periodic boundary conditions.

In order to shed more light on the role of the lifetime of the bubbles, we calculate a distribution function for the average bubble duration (τ_{ABD}):

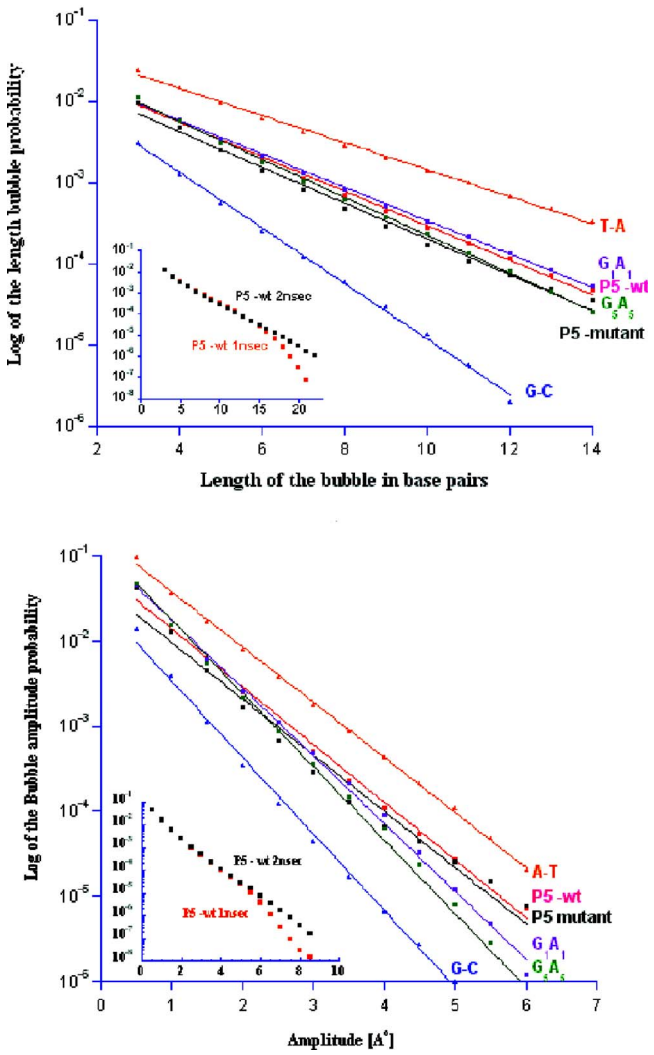

FIG. 5. (Color online) Bubble length probability (log scale) $\Sigma_{n=-46}^{23}\Sigma_{tr=1.5\ \text{Å}}^{\infty} P_n(l,tr)$ (upper) and bubble amplitude probability (log scale) $\Sigma_{n=-46}^{23}\Sigma_{l=3bp}^{\infty} P_n(l,tr)$ (lower). The insets compare these for AAV P5 at 1 nsec and 2 nsec simulation times.

$$\tau_{\text{ABD}}(n,l,tr) = \left\langle \frac{\sum_{q_n^k=1}^{q_n^{k_{max}}(l,tr)} \Delta t[q_n^k(l,tr)]}{\sum_{q_n^k=1}^{q_n^{k_{max}}(l,tr)} q_n^k(l,tr)} \right\rangle_M , \qquad (3)$$

where the denominator is the total number of bubbles, with strand separation tr, in the kth simulation, spanning l base pairs beginning at the nth base pair. It is important to emphasize that the information contained in the τ_{ABD} cannot be accessed through any thermodynamic considerations. In Fig. 4 we show the quantity $\Sigma_{tr=4.5\ \text{Å}}^{\infty}$-$\tau_{\text{ABD}}(n,l,tr)$ for the AAV P5 sequence as well as for the mutated AAV P5 sequence. The immediate observation is that the wild version of the AVV P5 sequence overall supports bubbles of significantly longer duration than the mutated version. This is particularly true for

bubbles of large strand separation. As documented by Fig. 3, the mutated AAV P5 certainly supports a number of large bubbles but their duration is significantly shorter. Also, Fig. 4 shows that the region around base pair +1 in the wild sequence supports large-amplitude, long-lived bubbles, a feature that is completely absent in the mutated sequence.

To compare the probability for bubble existence for all simulated sequences, we show in Fig. 5 the quantities $\Sigma_{n=-46}^{23}\Sigma_{tr=1.5\ \text{Å}}^{\infty}P_n(l,tr)$ (upper) and $\Sigma_{n=-46}^{23}\Sigma_{l=3bp}^{\infty}P_n(l,tr)$ (lower). In the two plots we show results for homogeneous A-T and G-C sequences, together with AAV P5 and its mutated version. Also shown are results for the two periodic sequences (G_1A_1) and (G_5A_5). All probabilities decrease exponentially with size (amplitude and length), rendering large bubbles rare dynamical events. As is natural given the softer A-T potential, the probability for any bubbles is always largest in a homogeneous A-T sequence and lowest in a homogeneous G-C sequence. Comparing the results for homogeneous, periodic, and heterogeneous sequences, it is clear that the bubble probability depends very little on sequence, but mainly on the AT and GC content. Finally, we observe that the periodic (G_5A_5) sequence has less probability for longer bubbles in comparison with the sequence with smaller period (G_1A_1) and with the heterogeneous AAV P5 sequence, confirming that the GC "barriers" and their impedance role is restricting energy flow in the sequence.

In the case of the bubble amplitude probability, there is a strong dependence on the actual sequence. The heterogeneous AAV P5 sequence sustains high-amplitude bubbles significantly better than the periodic sequences (G_1A_1) and (G_5A_5) with the same AT content. Even the mutated AAV P5 sequences with slightly less AT content than the periodic sequences (G_1A_1) and (G_5A_5) is more probable to sustain bubbles with amplitudes over 4 Å. Therefore the amplitude of the bubbles is sensitive to the exact sequence. In the heterogeneous sequences the probability for bubble with high amplitudes is larger than in the periodic sequences with the same or even a little less AT content. This is consistent with a recent demonstration [17] of melting temperatures being sensitive to intrasequence correlation rather than being simply determined by AT and GC content.

The insets of Fig. 5 compare the results of the 1-nsec and 2-nsec simulations for the AAV P5 sequence. Since the results are equivalent up to amplitudes larger that 6 Å and bubble lengths larger than 14 base pairs, we conclude that in this sequence finite-time effects in Langevin simulations exist only beyond these amplitudes and lengths.

In summary we have performed Langevin simulations of the PBD model of DNA and confirmed earlier results regarding the sequence dependence of bubble formation in agreement with results obtained on a purely thermodynamic basis. However, we find that the dynamics more sharply delineates the regions active for thermal strand separation because the life times of bubbles are directly accounted for. We find that the probability for larger bubbles (lengths and amplitudes) is higher for heterogeneous than for periodic sequences with the same A-T content. The important role of the length of the G-C "barriers" for bubble existence was identified. We find that the bubbles with maximum duration begin their existence at biologically significant sites and that these bubble initiation sites are different for bubbles with different amplitudes. Finally, we found a striking sensitivity of the bubble life time on sequence. Therefore we suggest that DNA's ability to sustain bubbles in some regions is a result of competition between length scales arising from the nonlinearity and the sequence heterogeneity and that this competition sensitively controls the bubble life times. Since specific biological functions are likely to be aided by long-lived openings of specific sizes, this information regarding size-lifetime relationships is directly relevant.

B.S.A. is grateful to Professor P. Littlewood for discussions. Research at Los Alamos is performed under Contract No. W-7405-ENG-36 for the U.S. DOE.

[1] B. F. Volkman *et al.*, Science **291**, 2429 (2001).
[2] E. Z. Eisenmesser *et al.*, Science **295**, 1520 (2002).
[3] M. Tomschik *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 3278 (2005).
[4] C. H. Choi, G. Kalosakas, K. Ø. Rasmussen, M. Hiromura, A. R. Bishop, and A. Usheva, Nucleic Acids Res. **32**, 1584 (2004); G. Kalosakas, K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva, Europhys. Lett. **68**, 127 (2004).
[5] K. B. Blagoev, B. S. Alexandrov, E. H. Goodwin, and A. R. Bishop, DNA Repair **5**, 863 (2006).
[6] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[7] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, R44 (1993).
[8] T. S. van Erp *et al.*, Phys. Rev. Lett. **95**, 218104 (2005); See also C. H. Choi, A. Usheva, G. Kalosakas, K. Ø. Rasmussen, and A. R. Bishop, *ibid.* **96**, 239801 (2006).
[9] Z. Rapti *et al.*, Europhys. Lett. **74**, 540 (2006).
[10] G. Alta-Bonnet, A. Libchaber, and O. Krichevsky, Phys. Rev. Lett. **90**, 128101 (2003).
[11] γ=0.05 ps^{-1} (γ=0.005 ps^{-1}) during (after) preheating.
[12] A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585 (1998).
[13] S. Ares, N. K. Voulgarakis, K. Ø. Rasmussen, and A. R. Bishop, Phys. Rev. Lett. **94**, 035504 (2005).
[14] N. K. Voulgarakis, A. Redondo, A. R. Bishop, and K. Ø. Rasmussen, Phys. Rev. Lett. **96**, 248101 (2006).
[15] Comparing trajectories resulting from two almost equal initial conditions we observed exponential growth of the differences between all the coordinates indicating the existence of at least one positive Liapunov exponent consistent with the analytical calculation in H. Qasmi, J. Barre, and T. Dauxois, e-print cond-mat/0407662.
[16] A. Sanches and A. Bishop, SIAM Rev. **40**, 579 (1988).
[17] G. Weber *et al.*, Nat. Phys. **2**, 55 (2006).