Healing length and bubble formation in DNA

Z. Rapti,[1] A. Smerzi,[2,3] K. Ø. Rasmussen,[2] and A. R. Bishop[2]
[1]*Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA and Department of Mathematics, University of Illinois at Urbana-Champaign, 1409 West Green Street, Urbana, Illinois 61801, USA*
[2]*Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*CNR-BEC-INFM, Università di Trento, I-38050 Povo, Italy*

C. H. Choi and A. Usheva
Endocrinology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, 99 Brookline Avenue, Boston, Massachusetts 02215, USA
(Received 30 December 2005; published 4 May 2006)

It has been suggested that thermally induced separations ("bubbles") of the DNA double-strand may play a role in the initiation of gene transcription, and an accurate understanding of the sequence dependence of thermal strand separation is therefore desirable. Based on the Peyrard-Bishop-Dauxois model, we show here that the bubble forming ability of DNA can be quantified in terms of a healing length $L(n)$, defined as the length (number of base-pairs) over which a base-pair defect affects bubbles involving n consecutive base-pairs. The probability for a bubble of size n is demonstrated to be proportional to the number of adenine-thymine base-pairs found within this length. The method for calculating bubble probabilities in a given sequence derived from this notion requires several order of magnitude less numerical effort than direct evaluation.

DOI: 10.1103/PhysRevE.73.051902 PACS number(s): 87.14.Gg, 63.20.Pw

I. INTRODUCTION

Local separation of double-stranded DNA into single-stranded DNA is fundamental to transcription and other important intracellular processes in living organisms. In equilibrium, DNA will locally denaturate when the free energy of the separated single-stranded DNA is less than that of the double-stranded DNA. Because of the larger entropy of the flexible single-strand, the much more rigid double-strand can be thermally destabilized locally to form temporary "bubbles" in the molecule even at physiological temperatures [1]. Considering this entropic effect together with the inherent energetic heterogeneity—guanine-cytosine (GC) base-pairs are 25% more strongly bound than the adenine-thymine (AT) bases—of a DNA sequence, it is plausible that certain regions (subsequences) are more prone to such thermal destabilization than others. In fact, recent work [2] demonstrates not only that such a phenomenon exists but more importantly that the location of these large bubble openings in a variety of DNA sequences coincide with sites active during transcription events. This discovery led to the hypotheses that thermal bubbles may assist transcription initiation, which otherwise has been thought to be entirely controlled by the action of site-specific enzymes. If confirmed, this represent a significant advance in the understanding of the relationship between local conformation and function in biomolecules. While there is no guarantee that this mechanism applies to all transcription initiation events, the present agreement is very encouraging. The agreement is based on the Peyrard-Bishop-Dauxois (PBD) model [3], which contains some essential basic ingredients—local constraints (nonlinearity), base-pair sequence (colored disorder), and entropy (temperature). The equilibrium thermodynamic properties of the model were numerically calculated from the partition function using the transfer integral operator method (TIO). (A complementary direct numerical evaluation of the partition function has been reported in Ref. [5].) This allows the precise evaluation of probabilities of bubbles as a function of temperature, location in a given base-pair sequence, and bubble size. In recent work [4], we reported that the probabilities of finding bubbles extending over n sites do not depend on specific DNA subsequences. Rather, such probabilities depend on the density of weakly bound AT base-pairs within a sequence of $L(n)$ base-pairs. Here we show that this characteristic length is simply related to the characteristic distance away from an AT base-pair—considered as a defect placed in a homogeneous GC-sequence—where the probability values of the base-pairs return to the GC bulk-value. Lastly, based on this concept of effective density approximation, we examine five different human promoter sequences, and demonstrate the striking agreement in the predictions from the two methods.

II. THE PBD MODEL AND THE TIO METHOD

Our characterization of a sequence's ability to form and sustain large local strand separations is based on the calculation of the probabilities $P_k(s)$ that starting at site s, at least k consecutive base-pairs are separated by a distance greater than t. Here, k ranges from 1 to 10, s ranges from base-pair 1 to base-pair $N-k+1$, where N is the length (in base-pairs) of the double strand, and t has the value 1.5 Å.

To describe our system—double-stranded DNA with a given length and composition—we use the PBD model. For each base-pair, this model includes one degree of freedom y_n, which is the transverse stretching of the hydrogen bonds between complementary bases. In the PBD model, the potential energy of a double strand of DNA of length N is

$$E = \sum_{n=1}^{N} [V(y_n) + W(y_n, y_{n-1})] = \sum_{n=1}^{N} \mathcal{E}(y_n, y_{n-1}). \quad (1)$$

Here $V(y_n) = D_n(e^{-a_n y_n} - 1)^2$ represents the hydrogen bonds between complementary bases; $W(y_n, y_{n-1}) = \frac{k}{2}(1 + \rho e^{-b(y_n + y_{n-1})})(y_n - y_{n-1})^2$ is the nearest-neighbor coupling that represents the (nonlinear) stacking interaction between adjacent base-pairs: it is comprised of a harmonic coupling with a state-dependent coupling constant effectively modeling the change in stiffness as the double strand is opened (i.e., entropic effects).

The heterogeneity of the sequence is incorporated by assigning different values to the parameters of the Morse potential, depending on the base-pair type. For simplicity, however, we use the same value for the constant k along each strand. We use parameter values (given in Refs. [6,9]) chosen to reproduce a variety of experimentally observed thermodynamic properties.

The equilibrium thermodynamic properties of the PBD model can be calculated from the partition function

$$\mathcal{Z} = \int \prod_{n=1}^{N} dy_n e^{-\beta \mathcal{E}(y_n, y_{n-1})} = \int \prod_{n=s}^{s+k-1} dy_n Z_k(s) e^{-\beta \mathcal{E}(y_n, y_{n-1})}, \quad (2)$$

where we have introduced the notation

$$Z_k(s) = \int \prod_{n \neq s, \ldots, s+k-1}^{N} dy_n e^{-\beta \mathcal{E}(y_n, y_{n-1})}$$

and $\beta = (k_B T)^{-1}$ is the inverse temperature. In order to evaluate the partition function (2) using the TIO method, we first symmetrize $e^{-\beta \mathcal{E}(x,y)}$ by introducing [7]

$$S(x,y) = \exp\left(-\frac{\beta}{2}[V(x) + V(y) + 2W(x,y)] \right) = S(y,x).$$

Here the second equality holds only when x and y correspond to base-pairs of the same kind. Using Eq. (2), the expression for $Z_k(s)$ is rewritten as

$$Z_k(s) = \int \left(\prod_{n \neq s, \ldots, s+k-1}^{N} dy_n S(y_n, y_{n-1}) \right) \times dy_1 e^{-(\beta/2)V(y_1)} e^{-(\beta/2)V(y_N)}, \quad (3)$$

where open boundary conditions at $n=1$ and $n=N$ have been used. To proceed, a Fredholm integral equations with a real symmetric kernel

$$\int dy S(x,y) \phi(y) = \lambda \phi(x) \quad (4)$$

must be solved separately for the AT and for the GC base-pairs.

Since the eigenvalues are orthonormal and the eigenfunctions form a complete basis, Eq. (4) can be used sequentially to replace all integrals by matrix multiplications in Eq. (3). Unlike in Ref. [8] where the kernels $S(x,y)$ were expanded in terms of orthonormal bases, here we choose to use Eq. (4) iteratively. In this way, we reduce the number of integral equations that need to be solved from four to two, and at the same time the matrices that need to be multiplied are lower dimensional. Whenever the sequence heterogeneity results in a nonsymmetric $S(x,y)$, Eq. (4) cannot be used and we resort to a symmetrization technique, based on successive introduction of auxiliary integration variables, as explained in Ref. [10].

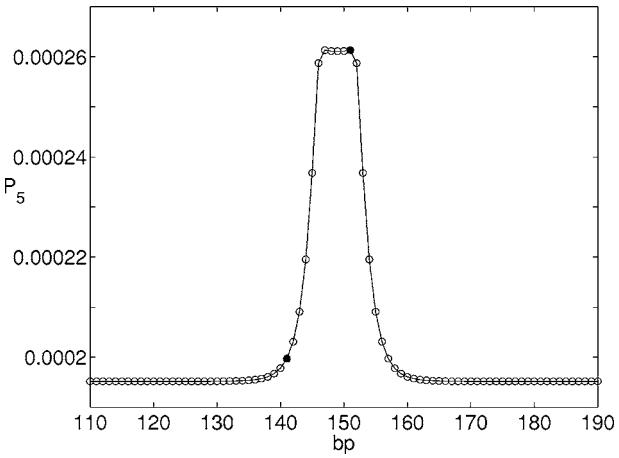

FIG. 1. The probability for the creation of a bubble of size five base pairs. The isolated AT base pair embedded in a sequence of GC base pairs at $n=151$ is shown by a filled black circle. A second black circle is located at base pair $151 - L(5) = 141$. The relative difference $\frac{P_5(141) - P_5(110)}{P_5(110)} = 0.0232$.

Using Eqs. (2) and (3), we evaluate the probabilities $P_k(s)$ as

$$P_k(s) = \mathcal{Z}^{-1} \int_t^{\infty} \prod_{n=s}^{s+k-1} dy_n Z_k(s) e^{-\beta \mathcal{E}(y_n, y_{n-1})}. \quad (5)$$

III. LENGTH SCALES AND EFFECTIVE DENSITY APPROXIMATION

In Ref. [4], we suggested that the probabilities of finding bubbles extending over n base-pairs to a very good approximation are proportional to the density of weakly bound AT base-pairs within a region of length $L(n)$. We choose the term effective density approximation (EDA) for the described approximation. The lengths $L(n)$ were obtained from numerical transfer integral calculations of the bubble probabilities of several simple (but experimentally realizable) sequences. The AT density profiles were therefore compared with the exact probabilities for thermal activation of bubbles of sizes $n=1$ and 5 of a wild and a mutant version of the adeno associated viral P5 promoter. Although the agreement was excellent, no physical explanation for the origin of these characteristic lengths was provided nor were they connected to any intrinsic length of the PBD model. However, since they appear prominently in the formation of DNA bubbles, it is important to investigate both of these questions.

In Figs. 1–3, we consider a sequence composed of 150 GC+1 AT+150 GC. In other words, we place a defect

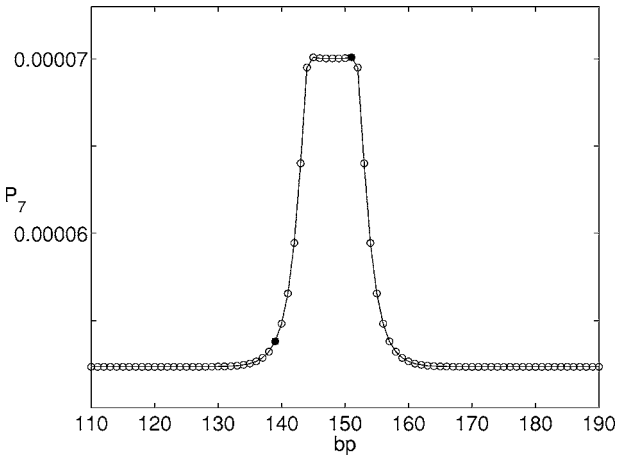

FIG. 2. The probability for the creation of a bubble spanning seven base pairs. The black circle at base pair 151 represents the defect. The second black circle is located at base pair $151-L(7)=139$. The relative difference $\frac{P_7(139)-P_7(110)}{P_7(110)}=0.0279$.

(AT instead of GC) at the site $l_0=151$. This defect is 150 base-pairs away from the two ends of the sequence in order to eliminate boundary effects.

AT base-pairs have a smaller bonding energy to GC base-pairs. Therefore, the AT defect weakens a number of GC base-pairs around it and increases the opening probability. Sufficiently away from the defect, the opening probability regains the bulk value of a homogeneous GC sequence at the given temperature, given threshold, and given bubble size. Our claim is that the characteristic length $L(n)$ is the distance over which the presence of AT base-pairs affects the GC base-pairs. This can be quantified by calculating the relative fluctuation

$$\frac{P_n(l_0-L(n))-P_n(110)}{P_n(110)}=\alpha, \tag{6}$$

where $l_0-L(n)$ is the base-pair obtained by counting $L(n)$ downstream from the position of the defect, see Figs. 1–3

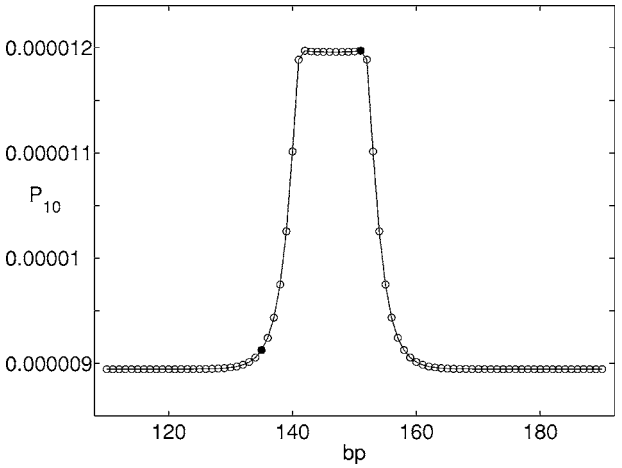

FIG. 3. The probability for the creation of a bubble spanning 10 base pairs. As before, the defect is represented by a solid black circle, and a second one is located at base pair $151-L(10)=135$. The relative difference $\frac{P_{10}(135)-P_{10}(110)}{P(110)}=0.0203$.

and at site 110 we assume that the bulk value has been regained. The remarkable finding is that with the choice of $L(n)$ considered in our previous work [4] (obtained independently by fitting the full numerical TIO calculations of the bubble formation probabilities of different simple sequences), we obtain from Eq. (6) $\alpha\simeq 2.5\%$, independently from the size of the bubble and the temperature of the DNA sequence. This can be seen in Figs. 1–3: the circle at base-pair $151=l_0$ is the AT defect, while the circles at base-pair 141, 139, 135 are the positions of the base pair at $l_0-L(n)$. We can therefore reverse the perspective and *define* the characteristic length as the one given by Eq. (6), with $\alpha\simeq 2.5\%$. This is important for practical applications, since it gives a simple criterion to estimate bubble probabilities for arbitrary bubble sizes and DNA temperatures (and arbitrary PBD inter base-pair interaction parameters), but it also immediately suggests a simple physical explanation for $L(n)$.

We parametrize the decay of the probability values as a function of the downstream distance from the AT defect by

$$P_n(l)=A_n+B_n\exp\left[-\frac{l_0-n+1-l}{\xi_n}\right], \tag{7}$$

where $P_n(l)$ is the probability for finding a bubble of size n located at the site l, A_n is the bulk value of the homogeneous GC sequence, namely the value of $P_n(l)$ calculated far away from the defect, and A_n+B_n is the value of the probability at the site l_0-n+1, which is the same as the probability value of the defect site l_0. ξ_n is the healing length of the system, which is the characteristic length required for the effects of the perturbation to die out, which depends on the size n of the bubble, temperature of the DNA, and the parameters of the PBD model. Replacing Eq. (7) in Eq. (6), we obtain the relation $L(n)=n-1+\xi_n\ln\left(\frac{B_n}{\alpha A_n}\right)$, where $B_n/A_n=[P_n(l_0)-P_n(\text{bulk})]/P_n(\text{bulk})\simeq 0.34$. We emphasize that both B_n/A_n and α are independent of the size n of the bubble. It follows that there is a simple linear relation between the healing and characteristic lengths,

$$L(n)=n-1+2.6\xi_n. \tag{8}$$

The healing length can now easily be calculated by inverting Eq. (7) (with an arbitrary value of l) as a function of the bubble size and temperature with a homogeneous GC sequence plus a single defect. From this, we can calculate the value of $L(n)$ and, therefore, estimate the probability for the creation of bubbles for arbitrary DNA sequences at any temperature. For instance, for bubbles of size $n=7$, we obtain $L(7)=12$, while for bubbles of size $n=10$ we have $L(10)=16$ at $T=300$ K and PBD parameters as in [9].

In order to examine how the values of the parameters in the PBD model affect $L(n)$, we set $\rho=0$ and repeat the calculation of $L(10)$. Since, when ρ decreases, so does the "cooperativity" of the base base pairs, one would expect to observe a drop in the $L(10)$ value. This is indeed the case: $L(10)=14$, while for $\rho=2$ the value is 16.

We will show in the next section how this approach compares with exact transfer integral operator calculations of the statistical properties of the PBD model.

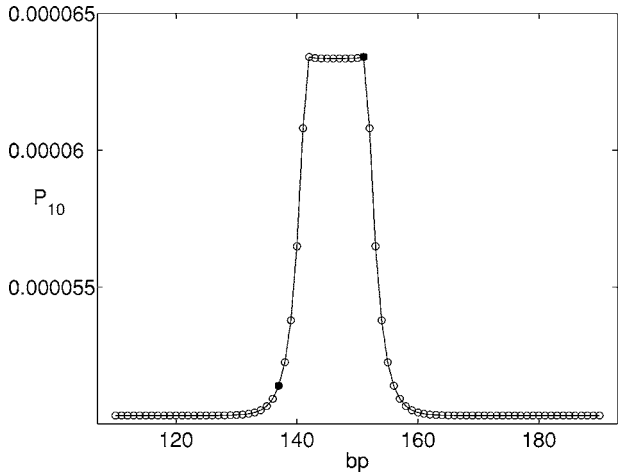

FIG. 4. In this figure, we show the probability for the creation of a bubble spanning 10 base pairs, when the coupling constant $\rho=0$. As is indicated by the relative difference $\frac{P_{10}(137)-P_{10}(110)}{P(110)}=0.0216$, now $L(10)=14$.

We conclude this section by noting that Figs. 1–3 show no symmetry with respect to the defect. While the defect is always at base pair 151, the symmetry is with respect to base pair 149 in the P_5 case, 148 in the P_7 case, and the axis that separates base pairs 146-147 in the P_{10} case. Another feature is the existence of a second local maximum with the same value as $P_n(151)$, and a slight drop in the probability values in the middle of the peak. We notice that the two maxima are located at sites l_0 and l_0-n+1. This suggests that a bubble with a defect at its boundary has a higher probability to form: in the $P_n(l_0-n+1)$ case the defect is at the end of the bubble, while in the $P_n(l_0)$ case it is at the beginning of the bubble. Also, the probability drops in the middle of the peak because the bubble there contains a defect that is trapped within GC base pairs, and it turns out that the probability of formation of a bubble of this kind is smaller (see Fig. 4).

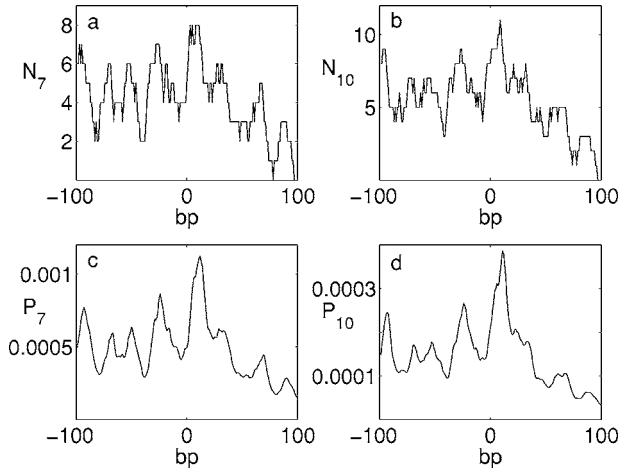

FIG. 5. Effective density profiles for 7- and 10-site long bubbles (a,b) and probability profiles calculated with the transfer integral approach (c,d). The sequence is the cox 8 promoter.

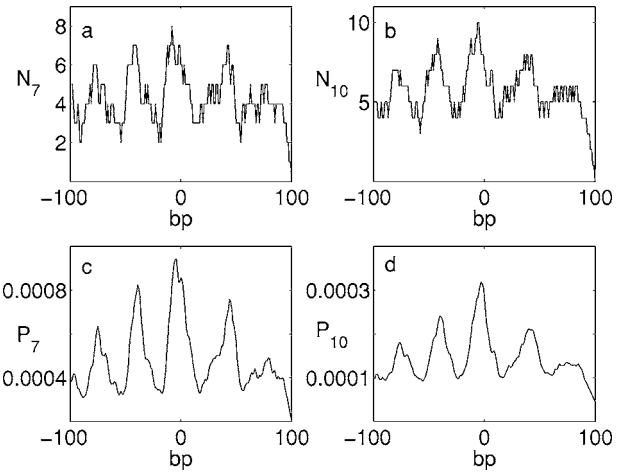

FIG. 6. Effective density profiles for 7- and 10-site long bubbles (a,b) and probability profiles calculated with the transfer integral approach, (c,d). The sequence is the cox 11 promoter.

IV. COMPARISON OF THE EDA AND TIO METHOD

We now compare the probability profiles obtained from the effective density approach with the characteristic length $L(n)$ calculated as discussed in the previous section, with exact results obtained with the TIO method. We consider five different human genome subsequences, and compare the calculations for the probability of formation of bubbles of sizes $n=7$ and 10.

In the panels (a,b) of Figs. 5–9 we plot, as a function of the base pair site, the numbers N_7 and N_{10} (which are proportional to the probability of formation of bubbles of sizes $n=7$ and 10, respectively) of AT base pairs calculated over a distance $L(7)=12$ (panel a) and $L(10)=16$ (panel b). These AT density profiles can be compared with the probability for the thermal creation of bubbles of seven, P_7, and ten, P_{10}, sites, panels (c, d). In all cases (and in several others not reported here), the resemblance in the main features of the respective profiles is striking. In particular, EDA correctly

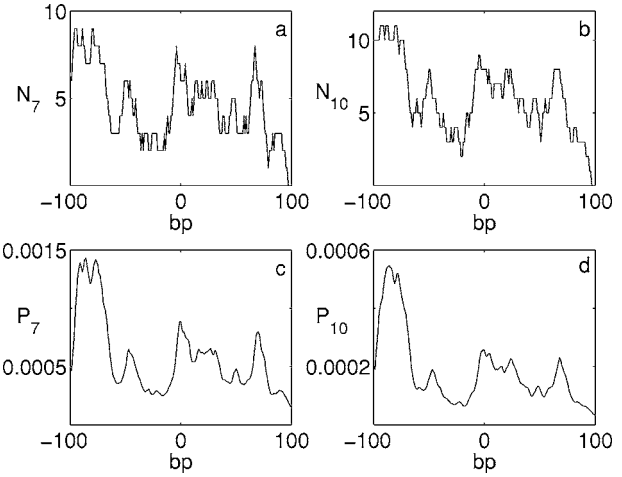

FIG. 7. Effective density profiles for 7- and 10-site long bubbles (a,b) and probability profiles calculated with the transfer integral approach (c,d). The sequence is the gtf2f2 promoter.

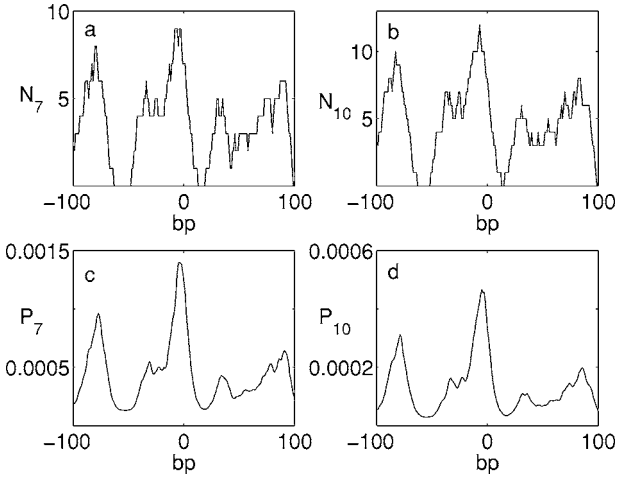

FIG. 8. Effective density profiles for 7- and 10-site long bubbles (a,b) and probability profiles calculated with the transfer integral approach (c,d). The sequence is the h33a promoter.

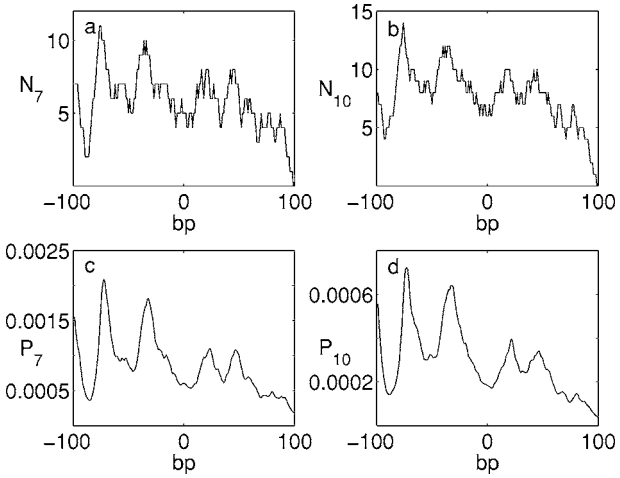

FIG. 9. Effective density profiles for 7- and 10-site long bubbles (a,b) and probability profiles calculated with the transfer integral approach (c,d). The sequence is the h3b promoter.

predicts the locations and relative weights of the probability peaks. The crucial point is that, while the profiles obtained with the EDA require only a few seconds, the full TIO method is very time consuming (of the order of several hours in the cases presented here). To fully appreciate this advantage, we note that with the EDA the entire human genome can be sequenced for bubble formation probabilities in a matter of minutes, while a statistical approach based on the calculation of the partition function is clearly impossible.

V. CONCLUSIONS

It has been suggested that the DNA transcription initiation sites may coincide with the location of large bubble openings. A thorough investigation of this hypothesis requires the statistical analysis of many DNA promoters within the PBD model. Such a task quickly becomes prohibitive when studying bubble-promoter correlations in a large number of cases. This problem has motivated the development of a simple alternative approach to calculate the bubble formation probabilities. We have found that these probabilities are proportional to the density of soft AT base-pairs calculated over a length that depends on the size of the bubble and the temperature. We have clarified the physical origin of this length and suggested a simple procedure for its calculation. The results of our effective density approach are in good agreement with exact calculations, but with the numerical effort reduced by several orders of magnitude. In this way, the entire human genome can now be analyzed, opening a unique possibility to understand the existence and nature of the correlations between thermally activated bubbles and promoters.

ACKNOWLEDGMENTS

Work at Los Alamos National Laboratory was supported by the U.S. Department of Energy under Contract No. W-7405-ENG-36.

[1] M. Gueron, M. Kochoyan, and J. L. Leroy, Nature (London) **328**, 89 (1987); M. Frank-Kamenetskii, *ibid.* **328**, 17 (1987).
[2] C. H. Choi, G. Kalosakas, K. Ø. Rasmussen, M. Hiromura, A. R. Bishop, and A. Usheva, Nucleic Acids Res. **32**, 1584 (2004).
[3] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989); T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. E **47**, R44 (1993).
[4] Z. Rapti, A. Smerzi, K. Ø. Rasmussen, A. R. Bishop, C. H. Choi, and A. Usheva, Europhys. Lett. **74**, 540 (2006).
[5] T. S. van Erp, S. Cuesta-Lopez, J.-G. Hagmann, and M. Peyrard, Phys. Rev. Lett. **95**, 218104 (2005); C. H. Choi, A. Usheva, G. Kalosakas, K. Ø. Rasmussen, and A. R. Bishop Phys. Rev. Lett. (to be published).
[6] A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585 (1998).
[7] T. Dauxois and M. Peyrard, Phys. Rev. E **51**, 4027 (1995).
[8] Y. L. Zhang, W.-M. Zheng, J.-X. Liu, and Y. Z. Chen, Phys. Rev. E **56**, 7100 (1997).
[9] The parameters were chosen in Ref. [6] to fit thermodynamic properties of DNA: $k=0.025$ eV/A^2, $\rho=2$, $\beta=0.35$ A^{-1} for the intersite coupling; for the Morse potential $D_{GC}=0.075$ eV, $a_{GC}=6.9$ A^{-1} for a GC base pair, $D_{AT}=0.05$ eV, $a_{AT}=4.2$ A^{-1} for the AT base pair.
[10] M. B. Fogel, Ph.D. thesis, Cornell University (1977).