The effect of sequence correlation on bubble statistics in double-stranded DNA

Jae-Hyung Jeon  
Department of Physics, Pohang University of Science and Technology, Pohang 790-784, South Korea

Pyeong Jun Park  
School of Liberal Arts, Chungju National University, Chungju 380-702, South Korea

Wokyung Sung  
Department of Physics and PCTP, Pohang University of Science and Technology, Pohang 790-784, South Korea

(Received 9 August 2006; accepted 11 September 2006; published online 24 October 2006)

DNA exists stably in the double-stranded structure at physiological temperatures, but base pairs are observed to unbind locally, giving way to bubbles (i.e., locally denatured states) due to thermal fluctuation. In this study, we consider the effect of sequence on the bubble statistics. On the basis of the Edwards equation description [W. Sung and J.-H. Jeons, Phys. Rev. E 69, 031902 (2004)], we develop a stochastic model incorporating the sequence randomness as a dichotomic noise, where the bubble and its size are identified as a returning random walk and its first passage time, respectively. By simulating the model Langevin equation, we obtain the bubble size distribution and show how it is affected by the sequence correlation. We find that the bubble size distribution of DNA with finite sequence correlation deviates from the Poland-Scheraga-type distribution. In particular, the formation of large bubbles is dramatically enhanced as sequence correlation length gets longer.

© 2006 American Institute of Physics. [DOI: 10.1063/1.2359724]

I. INTRODUCTION

DNA is a self-assembled heteropolymer of double-helical structure, which is composed of four different bases: adenine (A), thymine (T), cytosine (C), and guanine (G).1 Due to the complementary AT and GC pairs, the DNA duplex structure is thermodynamically stable at physiological conditions.1 However, it is observed to denature due to thermal fluctuation, as revealed by, e.g., UV absorption and fluorescence correlation spectroscopy experiments.2,3 A double-stranded (ds) DNA completely denatures into two single-stranded (ss) DNAs above a denaturation temperature (Tc) of about 350 K and undergoes local denaturation (i.e., bubble formation) even at room temperature.

Bubble formation is directly relevant to several biological processes. When the transcription process starts, DNA is partially opened for RNA polymerase to read the template. Currently, several theoretical calculations4 also showed a possibility that bubbles can dramatically enhance the probability of spontaneous DNA looping, which contributes significantly to gene regulation and stability of nucleosomes. On the other hand, bubble formation may lead to a corruption of genetic information by processes such as mutations or deletion of opened bases during the long lifetime of bubbles.3

The thermal denaturation of DNA has been extensively investigated, accompanied by several physical models.2 For example, the Poland-Scheraga (PS) model describes a ds-DNA as alternating sequences of bound segments and bubbles of various sizes.5 This model has shown that the denaturation transition is driven by the entropic gain of bubble formation and the transition nature is determined by the statistical weight of the bubble.5 The quantitative analysis of the equilibrium and dynamical properties of DNA have been done also by using the Peyrard-Bishop-Dauxois (PBD) model,6,7 where a dsDNA is assembled by the pairing interaction between complementary bases and by the stacking interaction between adjacent bases along the strand. For instance, melting curves obtained from the PBD model are in good agreement with experimental observations.8,9 The effects of helical structure,10 chain stiffness and excluded volume interaction,7,11-13 base stacking,14 and sequence heterogeneity15-17 on the denaturation have been studied on the basis of either the PS- or PBD-type model.

Recently, as experimental techniques advanced to probe local conformational change of DNA, statistics9,12,18-24 and dynamics3,25-29 of bubble formation have been investigated. Concerning bubble size distribution for homogeneous sequence, the PS model predicts that it follows an exponentially damped power-law decay below Tc and a power-law decay near Tc [see Eq. (4.3)]. Recently, the distribution was confirmed by several numerical and analytical studies,12,19,21,22 and was related to the nature of denaturation transition. Stochastic formulation was proposed to analytically and numerically calculate the bubble size distribution and correlation functions19,27-29.

The genome, in general, has complex characteristics of the base composition and the sequence correlation. Since a GC pair has three hydrogen bonds and an AT pair only two,
thermal opening behavior (and thus stability) of DNA should depend on the sequence. For instance, the denaturation temperature increases linearly with the percentage of GC pairs in a DNA, which in nature ranges from 22% to 73%. Interestingly, almost equal amounts of the GC and AT pairs exist in DNA of higher organisms, so the sequence correlation should be more relevant to the thermal stability of DNA.

Several studies have considered heterogeneous DNA sequences. Cule and Hwa considered a short-range-correlated random sequence and numerically found that the sequence randomness is irrelevant to \( T_c \), as well as to the critical behavior. Yeramian investigated the thermal opening behavior of chromosomes (~several 10^5 base pairs (bps)) consisting of multiple genes and found that the gene (coding) regions in a chromosome are thermally more stable than the intergenic (noncoding) regions. Choi et al. investigated the thermal opening probability of DNA promoter segments as two ideal strands pairwise interconnected by a bubble formation remaining to be studied systematically.

In accessing the sequence heterogeneity, one may note that, besides the base-pairing mentioned above, stacking interaction also contributes. It depends not only on the identity of neighboring nucleotides but also on their directions; for example, the interaction on 5’-G-A-3’ is different from that on 5’-A-G-3’. As a result, stacking interaction appears in ten different strengths. Simulations of the PBD model, which neglects this heterogeneity of stacking interactions while taking into account only the heterogeneity of hydrogen bonding between base pairs, have predicted bubble formation in excellent agreement with experiments. Presumably, it is because a bubble can be formed with a small amount of sliding of base, which may only partially disrupt the stacking. In this work, we focus on the effect of heterogeneity via the hydrogen bonding on bubbles.

To gain an analytical understanding of the effect, we use the simple version of the PBD model, which regards the dsDNA as two ideal strands pairwise interconnected by a Morse potential. Also, a very similar model to this was used by Lubensky and Nelson, who studied the effect of sequence randomness, treated as Gaussian and white noise, on denaturation induced by unzipping force. As we have shown earlier for homogeneous sequence, this model can be transformed to a stochastic model in which the statistics of bubble can be readily obtained. Further using this stochastic model in the present work, we incorporate the sequence randomness as a nonequilibrium noise with finite correlation length. By performing the Brownian dynamics simulation, we obtain the bubble size distribution and the cumulative bubble size distribution for various correlation lengths and temperatures. Our results show that due to the finite sequence correlation, bubble size distribution does not follow the well-known PS-type distribution. Moreover, the cumulative bubble size distribution indicates that large bubbles form more easily as the correlation length gets longer at all temperatures. We provide physical explanations for our simulation results and discuss their biological implications.

The paper is organized as follows. In Sec. II, we introduce the polymer Green’s function with sequence heterogeneity, and derive the equivalent Langevin equation for the interstrand distance of the dsDNA. In Sec. III, the procedure for the Brownian dynamics simulation is presented. In Sec. IV, our simulation results are given and discussed. Finally, a summary of the present study is given in Sec. V.

II. AN IDEAL CHAIN MODEL OF HETEROGENEOUS DNA

A. The polymer formalism

As mentioned above, we consider the PBD-type interaction energy that effectively describes the double strand conformation, which involves the effective strength of base pairing, the stacking interaction, and the sequence-dependent pairing potential; these are described by the Morse potentials,

\[
\mathcal{V} = \sum_{n=0}^{N-1} \left[ K (x_{n+1} - x_n)^2 + U_n (x_n) \right],
\]

(2.1)

where \( x_n \) is the interstrand distance of the \( n \)th base pair. The first term accounts for the simplified stacking interaction between adjacent bases. It determines the elastic property of DNA chain, and is responsible for the loop entropy factor in the PS model. The second term \( U_n(x) \) represents the sequence-dependent pairing potential; \( U_n(x) = U^*_n(x) \) for a GC pair and \( U_n(x) = U^\prime_n(x) \) for an AT pair, all of which are described by the Morse potentials,

\[
U^*(x) = D^* e^{-(x-x_m)/A} \left( e^{-(x-x_m)/A} - 2 \right).
\]

(2.2)

Here, \( A^{-1} \) is the potential width and \( x_m \) is the position of potential minimum. In this model, the difference of the binding strength between AT and GC pairs is given only by the effective strength of base pairing \( (D^*) \), which is determined by the hydrogen bonding, hydrophobic effect, counterion concentration, solvent pH, etc. We denote the DNA sequence as \( \{ \eta \} = \{ \eta(1), \eta(2), \ldots, \eta(N) \} \), where \( \eta(n) = +1 \) for GC pairs and \( -1 \) for AT pairs. Then, the pairing potential in a DNA with its sequence \( \{ \eta \} \) can be represented as

\[
U_n(x) = D (1 + p \eta(n)) \mathcal{U}(x),
\]

(2.3)

where \( \mathcal{U}(x) = e^{-A(x-x_m)/A} e^{-(A(x-x_m)/A) - 2} \) and \( D^* \) is rewritten as \( D (1 \pm p) = D^* \).

Now we introduce Green’s function \( G_\eta(x,n|x_0) \), the statistical weight that the \( n \)th base pair has a separation \( x \) provided the first one has a separation \( x_0 \) for a given sequence \( \{ \eta \} \). In the continuum limit where \( Nl \) is fixed, the Edwards equation for Green’s function is given by

\[
- \frac{\partial}{\partial x} G_\eta(x,n|x_0) = H G_\eta(x,n|x_0),
\]

\[
= \left[ H_0 + \eta(n) H' \right] G_\eta(x,n|x_0),
\]

\[
= \left[ - \frac{1}{6} \frac{d^2}{dx^2} + \beta D (1 + p \eta(n)) \mathcal{U}(x) \right] \times G_\eta(x,n|x_0),
\]

(2.4)
where \( \beta = 1/k_B T \) (\( k_B \) Boltzmann constant), \( \ell^2 = 3k_BT/2k_B \), and \( \ell/\sqrt{2} \) is the average intrastrand spacing between neighboring bases along its strand. Here, we denote the \( n \)-independent part of \( \mathcal{H} \) as \( \mathcal{H}_0 = -(\ell^2/6)(\partial^2/\partial x^2) + \beta D \mathcal{U}(x) \) and the \( n \)-dependent part of \( \mathcal{H} \) as \( \mathcal{H}' = \beta D p \mathcal{U}(x) \). Equation (2.4) is identical in form to the imaginary-time Schrödinger equation with time-dependent random potential that flips between \( \beta D(1+p) \mathcal{U}(x) \) and \( \beta D(1-p) \mathcal{U}(x) \). Equation (2.4) has been used to study the thermal \(^{6,15,19}\) and mechanical \(^{15}\) denaturation of DNA.

When \( \eta(n) \) is fixed to be either +1 (homogeneous GC-DNA) or -1 (homogeneous AT-DNA), the operator \( \mathcal{H} \) is \( n \) independent and the Green’s function can be written as

\[
G^\pm(x,n|x_0) = \sum_x \exp(-n\varepsilon_n^\pm) u_n^\pm(x) u_n^\pm(x_0),
\]

(2.5)

where \( u_n^\pm(x) \) and \( \varepsilon_n^\pm \) are the normalized eigenfunction and eigenvalue of \( \mathcal{H}_0 \pm \mathcal{H}' \), respectively. For large \( n \), \( G^\pm(x,n|x_0) \) can be dominated by the ground state term as

\[
G^\pm(x,n|x_0) \approx e^{-n\varepsilon_0^\pm} u_0^\pm(x) u_0^\pm(x_0),
\]

(2.6)

where

\[
|\varepsilon_0^\pm| = \frac{A^2\ell^2}{24} \left( \frac{T^\pm}{T} \right)^{1/2} - 1 \right)^2,
\]

(2.7)

and

\[
u_0^\pm(x) = e^{\pm \varepsilon_0^\pm} \exp\left[-\frac{1}{2} \left( \frac{T^\pm_x}{T} \right)^{1/2} e^{A(x-x_m)} - \frac{1}{2} A \left( \frac{T^\pm_x}{T} \right)^{1/2} (x-x_m) \right].
\]

(2.8)

Here, \( e^\pm \) is a normalization constant, and \( T^\pm_x = 24D^\pm/k_BT^\pm \) is the temperature at which the homogeneous GC- or AT-DNA completely denatures, respectively.

By convolution of \( G^+ \) and \( G^- \), in principle, \( G_n \) can be obtained for a given \( \{\eta(n)\} \). For instance, when \( \{\eta\} = \{+1, \ 1, -1, -1, -1, -1, -1, +1, +1, +1\} \), there are three domains of the same kind of base pairs and the Green’s function can be written as

\[
G_n(x,n|x_0) = \int dx dx_2 G^+(x,n_3|x_2)G^-(x_2,n_2|x_1)G^+(x_1,n_1|x_0).
\]

(2.9)

Here, \( n_1 = 2, \ n_2 = 4, \) and \( n_3 = 3 \) are the number of base pairs in each homogeneous domain and \( n = n_1 + n_2 + n_3 \). For a given DNA sequence, \( G_n \) itself can be expressed formally, and relevant physical quantities can be expressed in terms of \( G_n \) but taking the average over different sequences for those quantities is formidable. Moreover, it is not straightforward to use \( G_n \) in studying bubble size distribution. To overcome this difficulty, we employ the equivalent stochastic dynamics formulation in the following section.

**B. The stochastic formalism**

Our previous study \(^{19}\) for homogeneous DNA showed that the Edwards equation for either \( \eta(n) = \) or -1 can be transformed to the Fokker-Planck equation \(^{20}\) as

\[
\frac{\partial}{\partial t} P(x,n|x_0) = \mathcal{L}_{FP} P(x,n|x_0),
\]

(2.10)

where

\[
\mathcal{L}_{FP}^\pm = - e^{-\beta V^\pm(x)} H_{\pm}^x e^{\beta V^\pm(x)/2}
\]

(2.11)

and

\[
P^\pm(x,n|x_0) = e^{-\beta [V^\pm(x)-V^\pm(x_0)/2]} G^\pm(x,n|x_0).
\]

(2.12)

Here, the potential \( V^\pm(x) \) is related to \( U^\pm(x) \) via the following relation:

\[
U^\pm(x) + |\varepsilon_0^\pm| = D e^{\beta V^\pm(x)} \frac{\partial^2}{\partial x^2} e^{-\beta V^\pm(x)/2},
\]

(2.13)

with \( D = \ell^2/6 \). In Eq. (2.13), \( U^\pm \) is shifted by \( |\varepsilon_0^\pm| \) to ensure the existence of the stationary solution of the Fokker-Planck equation in Eq. (2.10) in the limit \( n \to \infty \). \( V^\pm(x) \) can be expressed explicitly in terms of \( u_0^\pm(x) \), the ground state eigenfunction for \( \mathcal{H}_0 \pm \mathcal{H}' \), as follows: \(^{19}\)

\[
\beta V^\pm(x) = -2 \log u_0^\pm(x) + \text{const},
\]

(2.14)

\[
= \left( \frac{T^\pm_x}{T} \right)^{1/2} e^{-A(x-x_m)} + \text{const} \left( \frac{T^\pm_x}{T} \right) - 1 (x-x_m) + \text{const}.
\]

(2.15)

Figure 1 shows a typical shape of \( \beta V^\pm(x) \); it has potential minima at \( x = x_m + A^{-1} \log[1 - \sqrt{T/T^\pm_x}]^{-1} \) below the critical temperature \( T^\pm \), and \( \beta V^\pm(x) \) approach to straight lines for large \( x \) as \( T \) goes to \( T^\pm \), respectively. The dotted line in Fig. 1 represents the mean potential \( \langle V \rangle = (V^+ + V^-)/2 \).
In Eq. (2.10), the interstrand distance $x(n)$ is effectively a diffusive process under the transformed potential $V^c(x)$ for $\eta=\pm 1$, respectively. Then the equivalent Langevin equation can be written as

$$\frac{d}{dn} x(n) = -\frac{dV(x)}{dx} + \frac{dV_d(x)}{dx} + \xi(n)$$

(2.16)

for $\eta=\pm 1$, respectively. Here, $\gamma=D/k_BT$, and $\xi(n)$ is a Gaussian and white noise characterized by

$$\langle \xi(n) \rangle = 0,$$

(2.17)

$$\langle \xi(n) \xi(n') \rangle = 2\gamma k_BT \delta(n-n').$$

(2.18)

$V(x)$ and $V_d(x)$ are defined as $V=(V^*+V^-)/2$ and $V_d=(V^*-V^-)/2$, so that the particle is actually subject to the GC or AT potential ($V^*$ or $V^-$) with $\eta=+1$ or $-1$, respectively. The Brownian dynamics simulation using Eq. (2.16) enables us to obtain the trajectory $x(n)$ and the bubble size distribution for homogeneous GC-or AT-DNA, as in our previous study. \cite{footnote}

Now we consider a heterogeneous DNA with its sequence defined by $\{\eta\}$. Rewriting the Langevin equation as

$$\frac{d}{dn} x(n) = -\frac{dV(x)}{dx} - \frac{dV_d(x)}{dx} \eta(n) + \xi(n),$$

(2.19)

we consider that the diffusive process is subject to the dichotomically fluctuating potential. That is, for a given sequence $\{\eta\}$, the fictitious Brownian particle moves under either $V^+(x)$ or $V^-(x)$ at each step. The Langevin equation, Eq. (2.19), provides us a direct scheme to simulate the trajectories, $x(n)$, as an alternative to the Green’s function of the time-dependent Schrödinger equation, Eq. (2.4), which is hard to implement. In our model, we consider the sequence randomness as quenched noise characterized by

$$\eta(n) = 0,$$

(2.20)

$$\eta(n) \eta(n') = e^{-|n-n'|/\xi_c},$$

(2.21)

where $\cdots$ means averaging over all possible realizations of the sequence $\{\eta(n)\}$ and $\xi_c$ is the sequence correlation length. There are two limiting cases. As $\xi_c$ goes to 0, the DNA sequence becomes completely disordered, i.e., $\eta(n) \eta(n') = \delta_{nn'}$, while as $\xi_c \to \infty$, the sequence becomes homogenous.

III. NUMERICAL PROCEDURE

We prepared a set of DNA sequences $\{\eta(n)\}$ with various sequence correlation lengths, i.e., $\xi_c=1, 10, 100, 1000$, and 50 000. For each value of $\xi_c$, we generated 100 different sequences $\{\eta(n)\}$, where $n=1,2,3,\ldots,50 000$. The generation of a sequence $\eta(n)$ was performed as follows. For the first base pair, $\eta(n=1) = +1$ or $-1$ with equal probability $1/2$. Then, for the following base pairs, $\eta(n)$ is chosen by the rule that flipping to the value different from the previous one occurs with the transition rate $W(1|-1)=W(-1|1)=1/2\xi_c$. This procedure generates a random DNA sequence $\{\eta(n)\}$, which looks like a telegraph signal with a finite correlation length $\xi_c$.

IV. RESULTS AND DISCUSSION

A. Denaturation temperature $T_c$

The denaturation temperature ($T_c$) can be obtained directly from the aforementioned Brownian motion analogy. The fictitious Brownian particle diffuses under the “time-dependent” potential that flips between $V^+(x)$ and $V^-(x)$ with a flipping frequency $(2\xi_c)^{-1}$ (see Fig. 1). When $T<T_c$, the particle is confined within the well of the mean potential $V(x)=(V^*+V^-)/2$, with occasional excursions that correspond to the bubbles. On the other hand, when $T>T_c$, the confinement effect disappears so that the particle moves away from the bound state, corresponding to the separation of the two strands. The $T_c$ naturally identified (and simultaneously confirmed) as the temperature at which the mean potential has a vanishing slope for sufficiently large $x$, is found to be

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig2.png}
\caption{The typical interstrand distance trajectories $[x(n)]$ obtained from the Langevin dynamics simulations for $\xi_c=10$ at $T=310$ and 350 K. The sequence $\eta(n)$ is depicted by the shaded (GC domain) and white (AT domain) regions.}
\end{figure}
FIG. 3. The bubble size distribution (BSD) at $T=310$ K for $\xi=1$, 10, 100, 1000, and 50 000. The linear slope (solid line) is $\Delta$ given by Eq. (4.3) and (4.4) for the reference DNA. The dashed and dotted curves represent Eq. (4.6) and Eq. (4.8), respectively.

The loop closure factor $n^{-3/2}$ in Eq. (4.3) is related to the entropy reduction to form a closed-loop conformation. The distribution in Eq. (4.3) is the well-known PS-type distribution of bubble size of homogeneous DNA. In our previous study, the energy gap is identified as the absolute value of the lowest eigenvalue of the corresponding Edwards equation, which reads

$$\Delta(T) = \frac{\Delta^2 - \left(\frac{T_c}{T}\right)^{1/2} - 1}{24}. \quad (4.4)$$

For both cases of GC and AT homogeneous DNA, the BSD should be given by Eq. (4.3) with the energy gap $\Delta^*(T)$ instead of $\Delta(T)$, respectively, which is explicitly given in Eq. (2.7). The inequality among the energy gaps holds as $\Delta^*(T) < \Delta(T) < \Delta^*(T)$ due to the difference in the base pair binding strength.

For the case of $\xi=1$ (i.e., the sequence is short-range correlated), the BSD almost coincides with the theoretical prediction (solid line) by Eqs. (4.3) and (4.4) for the reference DNA. This is because a weakly bound AT pair and a neighboring strongly bound GC pair can be viewed as two base pairs whose binding strength is given by $\Delta$ [see Fig. 6(a)]. This behavior is also valid at $T=330$ and 350 K, as

$$T_c = \frac{T_c^+ + T_c^-}{2} \left(1 + \frac{(1 - p^2)^{1/2}}{2}\right). \quad (4.1)$$

Using the typical values $T_c^+ = 340.0$ K and $T_c^- = 375.7$ K, we estimate that $p \approx 0.05$ and $T_c \approx 357.8$ K. One can see that $T_c$ depends on $p = (D^+ + D^-)/2$, i.e., weakly on the difference between the depths of $U^+$ and $U^-$, but does not depend on the sequence correlation length. The sequence-independence of $T_c$ is in accordance with the simulation result of Cule and Hwa for small $\xi$.15

B. Bubble size distribution: First passage time distribution

The time elapsed for the excursions of the fictitious Brownian particle (i.e., the first passage time for the returning random walk) is identified as the size of the bubble. Then the bubble size distribution (BSD), i.e., the distribution of the first passage time, can be obtained from the simulated trajectories. For a given random sequence $\{n(n)\}$ with a fixed $\xi$, we obtain the BSD, $P_\xi(n)$, by ten independent simulations. We then average $P_\xi(n)$ over the 100 independent sequences satisfying Eqs. (2.20) and (2.21) to obtain the BSD $P(n)$ as

$$P(n) = \int P_\xi(n), \quad (4.2)$$

which is the probability of finding a bubble of size $n$ in a heterogeneous dsDNA with a given sequence correlation length $\xi$. In Figs. 3–5 BSDs at $T=310, 330$, and 350 K are presented for various values of $\xi$.

Figure 3 shows the BSDs at $T=310$ K for $\xi=1$, 10, 100, 1000, and 50 000. Depicted for comparison is the BSD (solid line) for a homogeneous DNA whose transformed base pair potential is given by $V = (V^+ + V^-)/2$, called the reference DNA. For this homogeneous (reference) DNA, the BSD can be theoretically derived as

$$P(n) = \frac{N n^{3/2}}{N n^{3/2}}, \quad (4.3)$$

where $N$ is a normalization factor, and $\Delta(T)$ is the energy gap (i.e., the free energy difference) per base pair (in units of $k_B T$) between the bound and completely unbound base pairs.

FIG. 4. The bubble size distribution (BSD) at $T=330$ K for $\xi=1$, 10, 100, 1000, and 50 000. Equations (4.3), (4.6), and (4.8) are represented by the solid, dashed, and dotted lines, respectively.

The loop closure factor $n^{-3/2}$ in Eq. (4.3) is related to the entropy reduction to form a closed-loop conformation. The distribution in Eq. (4.3) is the well-known PS-type distribution of bubble size of homogeneous DNA. In our previous study, the energy gap is identified as the absolute value of the lowest eigenvalue of the corresponding Edwards equation, which reads

$$\Delta(T) = \frac{\Delta^2 - \left(\frac{T_c}{T}\right)^{1/2} - 1}{24}. \quad (4.4)$$

For both cases of GC and AT homogeneous DNA, the BSD should be given by Eq. (4.3) with the energy gap $\Delta^*(T)$ instead of $\Delta(T)$, respectively, which is explicitly given in Eq. (2.7). The inequality among the energy gaps holds as $\Delta^*(T) < \Delta(T) < \Delta^*(T)$ due to the difference in the base pair binding strength.

For the case of $\xi=1$ (i.e., the sequence is short-range correlated), the BSD almost coincides with the theoretical prediction (solid line) by Eqs. (4.3) and (4.4) for the reference DNA. This is because a weakly bound AT pair and a neighboring strongly bound GC pair can be viewed as two base pairs whose binding strength is given by $\Delta$ [see Fig. 6(a)]. This behavior is also valid at $T=330$ and 350 K, as

FIG. 5. The bubble size distribution (BSD) at $T=350$ K for $\xi=1$, 10, 50, 100, and 1000. The linear slope (solid line) is $\Delta$ given by Eq. (4.4) and the linear slope (dotted line) is $\Delta^*$ for GC homogeneous DNA.
shown in Figs. 4 and 5. Our result is consistent with previous work,\textsuperscript{15,17} which numerically showed that the denaturation behavior is unaffected by sequence randomness, in particular, near $T_c$, as far as $\xi_c \approx 1$.

As the correlation length becomes longer than $O(1)$, however, the sequence heterogeneity effect emerges, so that the BSD deviates from the PS-type distribution. A bubble smaller than $\xi_c$ exists within a single GC or AT domain, whereas a bubble larger than $\xi_c$ should exist over multiple domains, as illustrated in Fig. 2. For this reason, we classify the statistics of bubble size $n$ into two regimes of $n \leq \xi_c$ and $n \geq \xi_c$ in the following discussion.

For the bubbles smaller than the sequence correlation length, the simulation results can be explained as follows. When $1 \ll \xi_c \ll N$ (system size), as illustrated in Fig. 6(b), heterogeneous DNA can be viewed as alternating variable-length AT and GC domains of average size $\xi_c$. In this case, the number of base pairs at the domain boundaries is very few relative to the total number of base pairs, and each domain will allow bubbles with a size distribution of few relative to the total number of base pairs, and each domain boundaries is very few relative to the total number of base pairs, and each domain.

For the bubbles larger than the sequence correlation length, the simulation results can be explained as follows. We determine the total number of bubbles in all GC/AT domains, respectively. We determine $w^+\xi_c$ exists within a single GC or AT domain, whereas a bubble larger than $\xi_c$ should exist over multiple domains, as illustrated in Fig. 2. For this reason, we classify the statistics of bubble size $n$ into two regimes of $n \leq \xi_c$ and $n \geq \xi_c$ in the following discussion.

For the bubbles larger than the sequence correlation length, the simulation results can be explained as follows. We determine $w^+\xi_c$ exists within a single GC or AT domain, whereas a bubble larger than $\xi_c$ should exist over multiple domains, as illustrated in Fig. 2. For this reason, we classify the statistics of bubble size $n$ into two regimes of $n \leq \xi_c$ and $n \geq \xi_c$ in the following discussion.

As shown in Figs. 3 and 4, the theoretical prediction agrees well with the simulation results for the case that $\xi_c \approx N$. As in the case of $n \leq \xi_c$ discussed above, the BSD changes to

$$P(n) \sim \frac{e^{-n\Delta_{eff}}}{n^{3/2}},$$

(4.7)

where $\Delta_{eff} = \Delta_{eff}(\xi_c, T)$ is the effective energy gap determined from simulation results. In fact, $\Delta_{eff}$ is identified as the unbinding free energy per monomer of a random DNA with a sequence correlation length $\xi_c$. From our simulation results, we find that $\Delta_{eff}$ equals $\Delta$ for $\xi_c = 1$ and decreases to $\Delta^*$ as $\xi_c$ gets longer. For example, the BSD for $\xi_c = 10$, and 100 indeed follows the above explanation as can be seen in Figs. 3–5.

The BSD for $\xi_c$ equal to the system size ($N = 50,000$) in Figs. 3 and 4 represents the limiting case of $\xi_c \to \infty$ as the system size goes to infinity. In this case, each DNA sequence is composed of either all GC or AT pair. Hence, the ensemble-averaged BSD for each sequence follows the normalized distribution function for either GC or AT homogeneous DNA, $P^+(n) = \exp(-n\Delta^*)/n^{3/2}N^\alpha$, where $N^\alpha$ is the normalization factor for the GC/AT case, respectively. Therefore, the disorder-averaged BSD for this case can be written as

$$P(n) = \frac{1}{2}(P^+(n) + P^-(n)).$$

(4.8)

As shown in Figs. 3 and 4, the theoretical prediction agrees well with the simulation results for the case that $\xi_c \approx N$. As in the case of $n \leq \xi_c$ discussed above, the BSD changes to

$$P(n) \sim \frac{e^{-n\Delta_{eff}}}{n^{3/2}},$$

(4.7)

where $\Delta_{eff} = \Delta_{eff}(\xi_c, T)$ is the effective energy gap determined from simulation results. In fact, $\Delta_{eff}$ is identified as the unbinding free energy per monomer of a random DNA with a sequence correlation length $\xi_c$. From our simulation results, we find that $\Delta_{eff}$ equals $\Delta$ for $\xi_c = 1$ and decreases to $\Delta^*$ as $\xi_c$ gets longer. For example, the BSD for $\xi_c = 10$, and 100 indeed follows the above explanation as can be seen in Figs. 3–5.

The BSD for $\xi_c$ equal to the system size ($N = 50,000$) in Figs. 3 and 4 represents the limiting case of $\xi_c \to \infty$ as the system size goes to infinity. In this case, each DNA sequence is composed of either all GC or AT pair. Hence, the ensemble-averaged BSD for each sequence follows the normalized distribution function for either GC or AT homogeneous DNA, $P^+(n) = \exp(-n\Delta^*)/n^{3/2}N^\alpha$, where $N^\alpha$ is the normalization factor for the GC/AT case, respectively. Therefore, the disorder-averaged BSD for this case can be written as

$$P(n) = \frac{1}{2}(P^+(n) + P^-(n)).$$

(4.8)

As shown in Figs. 3 and 4, the theoretical prediction agrees well with the simulation results for the case that $\xi_c \approx N$. As in the case of $n \leq \xi_c$ discussed above, the BSD changes to

$$P(n) \sim \frac{e^{-n\Delta_{eff}}}{n^{3/2}},$$

(4.7)

where $\Delta_{eff} = \Delta_{eff}(\xi_c, T)$ is the effective energy gap determined from simulation results. In fact, $\Delta_{eff}$ is identified as the unbinding free energy per monomer of a random DNA with a sequence correlation length $\xi_c$. From our simulation results, we find that $\Delta_{eff}$ equals $\Delta$ for $\xi_c = 1$ and decreases to $\Delta^*$ as $\xi_c$ gets longer. For example, the BSD for $\xi_c = 10$, and 100 indeed follows the above explanation as can be seen in Figs. 3–5.

The BSD for $\xi_c$ equal to the system size ($N = 50,000$) in Figs. 3 and 4 represents the limiting case of $\xi_c \to \infty$ as the system size goes to infinity. In this case, each DNA sequence is composed of either all GC or AT pair. Hence, the ensemble-averaged BSD for each sequence follows the normalized distribution function for either GC or AT homogeneous DNA, $P^+(n) = \exp(-n\Delta^*)/n^{3/2}N^\alpha$, where $N^\alpha$ is the normalization factor for the GC/AT case, respectively. Therefore, the disorder-averaged BSD for this case can be written as

$$P(n) = \frac{1}{2}(P^+(n) + P^-(n)).$$

(4.8)

As shown in Figs. 3 and 4, the theoretical prediction agrees well with the simulation results for the case that $\xi_c \approx N$. As in the case of $n \leq \xi_c$ discussed above, the BSD changes to

$$P(n) \sim \frac{e^{-n\Delta_{eff}}}{n^{3/2}},$$

(4.7)

where $\Delta_{eff} = \Delta_{eff}(\xi_c, T)$ is the effective energy gap determined from simulation results. In fact, $\Delta_{eff}$ is identified as the unbinding free energy per monomer of a random DNA with a sequence correlation length $\xi_c$. From our simulation results, we find that $\Delta_{eff}$ equals $\Delta$ for $\xi_c = 1$ and decreases to $\Delta^*$ as $\xi_c$ gets longer. For example, the BSD for $\xi_c = 10$, and 100 indeed follows the above explanation as can be seen in Figs. 3–5.

The BSD for $\xi_c$ equal to the system size ($N = 50,000$) in Figs. 3 and 4 represents the limiting case of $\xi_c \to \infty$ as the system size goes to infinity. In this case, each DNA sequence is composed of either all GC or AT pair. Hence, the ensemble-averaged BSD for each sequence follows the normalized distribution function for either GC or AT homogeneous DNA, $P^+(n) = \exp(-n\Delta^*)/n^{3/2}N^\alpha$, where $N^\alpha$ is the normalization factor for the GC/AT case, respectively. Therefore, the disorder-averaged BSD for this case can be written as
One can observe that as \(T\) is increased from 100 to 1000, the thermal stability of DNA is hindered by the adjacent GC pairs which are quite stable at this temperature of the AT pairs is hindered by the adjacent GC pairs which are stable at 100, and 1000.

The thermal stability of DNA is hindered by the adjacent GC pairs which are quite stable at this temperature of the AT pairs is hindered by the adjacent GC pairs which are stable at 100, and 1000.

\[
Q(n) = \sum_{m=n}^{\infty} N_p(m), \tag{4.9}
\]

where \(N_p(m)\) is the total number of bubbles of size \(m\) found in our simulations for a given \(\xi_c\) and \(T\). Then, \(Q(n)\) is the total number of bubbles larger than or equal to \(n\) for the given \(\xi_c\) and \(T\).

Figure 7 shows the CBSD at physiological temperature \(T=310\) K for \(\xi_c=1, 10, 100, \) and \(1000\). It is observed that \(Q(n)\) for large values of \(\xi_c\) could be much larger than that for small values of \(\xi_c\). For instance, the number of bubbles larger than or equal to 200 bps is about 100 times larger at \(\xi_c=1000\) than at \(\xi_c=1\). This is because, for \(\xi_c=1\), the unbinding of the AT pairs is hindered by the adjacent GC pairs which are quite stable at this temperature [Fig. 6(a)] while the AT pairs in AT domains, for \(\xi_c \geq 1\), can easily unbind to form a large bubble [Fig. 6(b)]. This result implies that the probability of the formation of bubbles larger than or equal to a specified size is dramatically enhanced by the sequence correlation effect at physiological temperature.

Figure 8 shows the CBSD at \(T=350\) K (\(T_c < T < T_s\)). One can observe that as \(\xi_c\) gets longer, small bubbles are suppressed, while the formation of large bubbles is enhanced. This result can be understood as follows. At this temperature, most AT domains form domain-sized bubbles since they are almost all denatured. The AT-domain bubbles themselves, as well as the cooperation among those AT-domain bubbles, contribute significantly to the number of large bubbles, so that large bubbles are formed easily for larger \(\xi_c\). For DNAs with small \(\xi_c\), small bubbles of size order of \(\xi_c\) are mostly the AT-domain bubbles, whereas for DNAs with large \(\xi_c\), small bubbles can be formed only within the GC domain. The number of small bubbles in the former case scales as \(N/2\xi_c\), which is much larger than that in the latter case at this temperature. Therefore, formation of small bubbles is more favorable for the DNA with small \(\xi_c\), while formation of large bubbles is more favorable for the DNA with large \(\xi_c\).

V. SUMMARY AND CONCLUSION

In this paper, we have studied the effect of finite sequence correlation on the bubble statistics in a heterogeneous DNA. The stochastic model originally suggested for homogeneous DNA in our previous study was adapted to the case of heterogeneous DNA, and the corresponding Brownian dynamics simulation scheme was developed. The denaturation transition temperature was obtained analytically from our model, which depends weakly on the difference of the complementary base pair interaction strengths between AT and GC pairs but not on the sequence correlation length. From the simulated trajectories of the fictitious Brownian particle, we obtained the conformation of bubbles in a DNA with given sequence correlation, and analyzed the bubble size distribution (BSD) as well as the cumulative bubble size distribution (CBSD).

Our simulation results show that the BSD follows the PS-type distribution for the case of homogeneous DNA as well for heterogeneous DNA with short-range correlation. For those DNAs with finite sequence correlations, on the other hand, the BSD appears to be different from the PS-type distribution. This result can be explained physically from the observation that the DNA is composed of alternating various sequence domains of AT and GC pairs. Bubbles smaller than the correlation length reside in the homogeneous domains of AT and GC pairs, while bubbles larger than the correlation length exist over multiple domains. Applying the corresponding bubble statistics, the simulation results for the BSD were reasonably explained.

Near the denaturation transition temperature, we found that AT domains are almost all unbound and the BSD for bubbles smaller than the correlation length follow the PS-type distribution of the homogeneous GC-DNA, as expected. On the other hand, the BSD for bubbles larger than the correlation length follow the PS-type distribution with the energy gap replaced by an effective value. The effective energy gap becomes negligible near the transition temperature, so that the BSD seems to follow the power law.

Another important result of the present study is the CBSD defined by the number of bubbles larger than or equal to a specified size. As the correlation length gets longer, the probability of large bubble formation is found to increase dramatically. This observation is important in understanding the thermal stability of DNA. It is known that DNA is com-
posed of two distinct regions, i.e., coding and noncoding regions. In the coding region, the correlation length is quite short for compact accumulations of genetic information, whereas the noncoding region is long-range correlated. Our study suggests that due to the sequence correlation effect, the coding region is more thermally stable than the noncoding region. The DNA sequence itself may reduce the chance for the coding region to allow large bubbles, so that the gene is not corrupted by ambient chemical agents.

The present stochastic theory is constructed based on the early version of the PBD model, which describes the thermal and mechanical denaturation of DNA qualitatively well. To incorporate the stacking interaction in a quantitative way, the harmonic spring term in Eq. (2.1) should include the dependence on the interbase distance as suggested in the modified PBD model. In this case, the initiation of bubble formation will be harder and the denaturation transition will be more cooperative. As a consequence, the values of the effective free energy per base pair (Δ) and the loop closure exponent will be modified in the PS-type distribution [e.g., Eq. (4.3)] of bubble size, which serves an underlying basic distribution of bubble size for heterogeneous sequences in the present work. Nevertheless, since the generic form of this distribution is preserved, our results are expected to be valid qualitatively for the modified PBD model.

ACKNOWLEDGMENTS

This work was supported by APCTP (Focus Program), Brain Korea 21 Project, and MOST (KOSEF) through SBD-NCRC.