Exact method for numerically analyzing a model of local denaturation in superhelically stressed DNA

Richard M. Fye[1] and Craig J. Benham[2,*]
[1]*Sandia National Laboratories, MS-1111, P.O. Box 5800, Albuquerque, New Mexico 87185*
[2]*Department of Biomathematical Sciences, Mount Sinai School of Medicine, Box 1023, 1 Gustave Levy Place, New York, New York 10029*
(Received 9 December 1997)

Local denaturation, the separation at specific sites of the two strands comprising the DNA double helix, is one of the most fundamental processes in biology, required to allow the base sequence to be read both in DNA transcription and in replication. In living organisms this process can be mediated by enzymes which regulate the amount of superhelical stress imposed on the DNA. We present a numerically exact technique for analyzing a model of denaturation in superhelically stressed DNA. This approach is capable of predicting the locations and extents of transition in circular superhelical DNA molecules of kilobase lengths and specified base pair sequences. It can also be used for closed loops of DNA which are typically found *in vivo* to be kilobases long. The analytic method consists of an integration over the DNA twist degrees of freedom followed by the introduction of auxiliary variables to decouple the remaining degrees of freedom, which allows the use of the transfer matrix method. The algorithm implementing our technique requires $O(N^2)$ operations and $O(N)$ memory to analyze a DNA domain containing N base pairs. However, to analyze kilobase length DNA molecules it must be implemented in high precision floating point arithmetic. An accelerated algorithm is constructed by imposing an upper bound M on the number of base pairs that can simultaneously denature in a state. This accelerated algorithm requires $O(MN)$ operations, and has an analytically bounded error. Sample calculations show that it achieves high accuracy (greater than 15 decimal digits) with relatively small values of M ($M<0.05N$) for kilobase length molecules under physiologically relevant conditions. Calculations are performed on the superhelical pBR322 DNA sequence to test the accuracy of the method. With no free parameters in the model, the locations and extents of local denaturation predicted by this analysis are in quantitatively precise agreement with *in vitro* experimental measurements. Calculations performed on the fructose-1,6-bisphosphatase gene sequence from yeast show that this approach can also accurately treat *in vivo* denaturation. [S1063-651X(99)06003-1]

PACS number(s): 87.15.By, 05.90.+m

INTRODUCTION

Unconstrained linear DNA molecules in solution at physiological temperatures and ionic conditions adopt the well-known Watson-Crick *B*-form structure, a right handed double helix conformation. However, DNA can occur in several other conformations. The biologically most important alternate conformation is the locally denatured (i.e., strand-separated) state, in which the base pairing between the two strands of the *B*-form DNA duplex is locally disrupted. Because local denaturation is an essential step in both transcription and replication, the two central functions of DNA, its occurrence must be stringently controlled *in vivo*. One means of exerting this control involves topological regulation of the unwinding torsional stresses that are imposed on the DNA [1].

DNA within living systems is organized into topological domains, typically several kilobases in length, consisting either of circular molecules or of closed loops within larger molecules [2]. The topological constraint on a closed-loop domain is precisely equivalent to that on a circular molecule. In either case the linking number L of the domain is fixed. L is the number of times either strand of the DNA links through the closed circle formed by the other strand. Equivalently, L is the total number of turns either strand of the DNA makes about the central axis curve of the domain, counted according to sign, when that central axis curve is planar. The linking number L is a global topological invariant of the domain: so long as both strands of the DNA remain continuous, the value of L cannot change. However, enzymes can alter the linking numbers of domains by processes involving transient strand breakage and relinking [3]. *In vivo*, these enzymes typically act (often in concert with other processes) to decrease the linking number L below the value L_o characteristic of the unstressed *B*-form double helix [3,4]. The resulting (negative) linking difference $\alpha=L-L_o<0$ imposes untwisting torsional stresses on the DNA domain involved, placing it in a (negatively) superhelical state. The domain can accommodate this condition in two ways. First, the *B*-form helix can bend and twist, deformations that require energy. Second, local regions of the DNA domain can undergo conformational transitions, such as denaturation, that decrease the helicity of the sites involved. These transitions accommodate part of the imposed linking difference α, which allows the rest of the domain to relax by a corresponding amount. Denaturation will be energetically favored when the energy of deformation relieved by this partial relaxation exceeds the cost of the conformational transition [5]. The localization of denaturation at specific sites within a domain

*Author to whom correspondence should be addressed.

results from the sequence dependence of the denaturation energy. Under a wide variety of environmental conditions *AT* base pairs on average require less free energy to separate than do *GC* pairs, with significant modulation due to near neighbor effects [6–9]. Hence, sites of local denaturation tend to be concentrated at *AT*-rich regions within a negatively superhelical domain.

The global superhelical constraint of fixed linking number L effectively couples together the conformational states of all the base pairs within a DNA domain. A transition at any one site alters its helical twist, which changes the distribution of the linking difference α throughout the domain, and thereby alters the stresses experienced by all other base pairs. This effective global coupling can lead to qualitatively entirely different types of transition behaviors than occur in the thermal denaturation of unconstrained polymers [10,11]. In unstressed linear polymers undergoing thermal denaturation, the probability of transition of each monomer typically increases monotonically with temperature. However, the imposition of negative superhelical on a DNA domain can lead to much more complex transition behavior. For example, the probability of denaturation of individual base pairs need not increase monotonically with the denaturing constraint α. Instead, as this negative superhelicity becomes more extreme, denaturation at new sites may be coupled to reversions back to *B* form (i.e., rejoining) of sites that had been denatured [1]. Moreover, whereas local changes of base sequence have at most local effects on thermally driven transitions [12–14], in stress-induced denaturation small local sequence alterations can have global consequences. For example, deletion of a 16-bp (base pair) region in a 4-kb circular DNA completely changes the locations and extents of superhelical denaturation throughout the molecule [15,16].

Stress-induced local denaturation has been shown experimentally to be involved in several important biological processes. The unique replication origin in the *E. coli* genome contains a stress-destabilized site located at a specific position relative to other markers [15]. Sequences that have been modified in a way that preserves the susceptibility of this site to superhelical denaturation retain their *in vivo* activity, while mutations that either degrade this susceptibility or move the position of the denaturing region by as few as 50 base pairs destroy the function of the replication origin. No other sequence specificity is observed around this position. Stress-induced denaturation also plays several known roles in DNA transcription. For example, expression of the *c-myc* oncogene is regulated in part by the upstream far upstream sequence element (FUSE) region, which is denatured *in vivo* under conditions where this gene is transcriptionally active [17]. Transcription of *c-myc* requires binding of the FUSE binding protein (FBP) regulatory protein to the unpaired DNA strands at the FUSE site. In a second example, expression of the $ilvP_G$ operon in *E. coli* is enhanced by binding of integration host factor (IHF) to a site 90 base pairs upstream of the transcription start site. However, this enhancement only occurs when the negative superhelicity of the DNA is sufficient to drive denaturation of the region abutting the IHF binding site. Under these circumstances IHF binding forces this site to revert to *B* form, which causes the next most easily destabilized region, around the transcription start site, to denature [18]. Stress-induced denaturation also has been implicated in the termination of transcription [19]. Sequence alterations that degrade the susceptibility to denature within the 3′ terminal flank of the yeast FBP1 gene decrease its frequency of correct termination *in vivo*. Stress-denaturable sites are involved in the binding of DNA to other cellular structures. These include sites where the DNA attaches to the chromosomal matrix [20,21] as well as the centromere region, where the DNA binds to the cellular apparatus that separates the two copies of a chromosome at cell division [22].

Local denaturation is the most extreme form of stress-induced DNA duplex destabilization. However, less extreme forms also may be biologically important. A process requiring local separation of the duplex strands may be controlled by proteins that can contribute some energy to this unpairing event, but not enough to drive it unless the site involved is already marginally destabilized. Such sites may become active when stresses are imposed on the DNA that decrease the energy required for their local denaturation without necessarily causing complete opening. For this reason it is also important to understand how imposed stresses affect the incremental energy required to denature individual sites within a DNA sequence.

Several approximate methods have been developed to analyze conformational transitions in superhelical DNA molecules. The first theoretical treatment demonstrated that local denaturation can be energetically favored in molecules experiencing untwisting torsional stresses [23]. It performed a simple two-state analysis in which only the competition between the untransformed state and the single energetically most favored denatured state was considered. Thereafter, three approximate statistical mechanical methods were developed to treat this problem.

The first of these approaches was a modification of the standard method to analyze helix-coil transitions in unconstrained linear molecules [24,25]. DNA that is circular was analyzed as though it were linear, with one unopenable base pair added to each end to reduce end effects. Then an energy renormalization step was performed to account approximately for the effects of the superhelical constraint. This approach did not impose the correct topological condition on the DNA. It also incorrectly assumed that sites of denaturation were torsionally undeformable, so that no part of an imposed linking difference α could be absorbed by interstrand twisting at the denatured sites. Single stranded DNA actually is highly flexible, so that large amounts of the imposed linking difference can be absorbed by such twisting at little cost in energy [26]. These oversimplifications severely limited the accuracy and utility of this method. Recently this approach has been extended by developing a more sophisticated self-consistent renormalization technique [27].

A second approximate analytic method has been developed that imposes the correct topological condition on the DNA domain and includes the torsional deformability of the unpaired regions [5,16]. In this approach an energy threshold θ is specified, and all states of denaturation are found whose energies exceed that of the minimum energy state by no more than this threshold amount. The cumulative effect of all states whose energies do not satisfy this threshold condition is estimated by a density of states calculation. From this information an approximate partition function is constructed, and approximate ensemble average values of important pa-

rameters are calculated. These include both the denaturation probability and the destabilization energy of each base pair in the DNA sequence. The accuracy of this method increases as the energy threshold θ is raised, although the number of included states, and hence also the computation time, grow approximately exponentially. This technique has internal controls that allow the user to achieve a specified level of accuracy by setting the energy threshold appropriately. Sample calculations show that reasonable accuracy can usually be achieved using moderate thresholds, for which the algorithm implementing this approximate method executes efficiently.

Extensive calculations have shown that the predictions of this approximate method are in close quantitative agreement with experimental results [16]. Experimentally determined energy parameter values are used in these calculations, so they have no free parameters. However, comparisons with experiments show that they correctly predict the sites and extents of denaturation as functions of the imposed linking difference, as well as the substantial effects on transition behavior that can result from even modest base sequence modifications. The close accord with experiment achieved by this method has enabled its use to predict the stress-induced destabilization properties of DNA sequences for which experimental information is not available [1,28].

Despite its successes, this approximate method has several shortcomings. It treats denaturation as copolymeric, with one separation energy ascribed to *AT* base pairs and another to *GC* pairs. As presently constructed it cannot handle more detailed transition energies, such as arise from the influence of near neighbor base pair identities, chemical modification of bases, bound ligands, abasic sites, pyrimidine dimers, or other molecular lesions. All of these local alterations of DNA are known to occur *in vivo*, and all can have a variety of important biological effects [29–33]. This method would also be difficult to extend to analyze competitions between local denaturation and other types of transitions in topologically constrained DNA molecules. And for technical reasons it cannot handle cases where the low energy states contain four or more distinct sites of simultaneous denaturation, as can occur in very long molecules ($N > 15\,000$ bp) or at high temperatures.

The third method that was developed to analyze conformational transitions in superhelical DNA is a generalized Monte Carlo sampling technique [34]. This approach can treat some of the special cases that neither of the previously described methods could handle, such as high temperatures and very long molecules. Also, it can be easily extended to treat multiple competing transitions. However, it is difficult to determine the frequency of occurrence of high energy states accurately using Monte Carlo sampling, so this method can estimate the destabilization energies of only the most strongly destabilized base pairs. It is comparatively slow to execute, and its accuracy in calculating many quantities is often less than that achieved by alternative methods. For these reasons Monte Carlo is the method of choice only in cases where no alternative approach is feasible.

In this paper we present a numerically exact technique to calculate the equilibrium properties of the denaturation (strand separation) transition in circular superhelical DNA molecules of specified base sequence and kilobase length. This approach has several advantages over its predecessors. It imposes the correct topological constraint on the DNA domain, it can treat any dependence of base pair separation energies on base sequence, and it can handle many situations which other methods find intractable. These include large numbers of simultaneously open regions, high temperatures, near neighbor effects, and the presence of chemical modifications such as abasic sites, methylated bases, lesions, or adducts. It explicitly treats fluctuations of interstrand twisting within denatured regions. Some earlier treatments ignored this phenomenon entirely [24,25], while others assumed these torsional deformations occurred at a mechanical equilibrium configuration of minimum energy [5,16,23,27]. The exact method also can be extended to treat competitions between denaturation and other types of transitions. As the exact contributions from all states are included, the accuracy of the method is limited only by its computational implementation. We also discuss an accelerated approximate implementation of the algorithm with analytic error bounds which can provide a speed-up of about an order of magnitude while retaining high accuracy.

DERIVATION OF THE METHOD

''Effective Hamiltonian'' and free energy considerations

We consider a closed circular DNA molecule containing N base pairs, on which a linking difference α has been imposed. Each base pair is regarded as being susceptible to transition to an alternative secondary structure (i.e., different helical structure and/or separation of the strands of the duplex). Here the alternative secondary structure is assumed to be local denaturation (strand separation), although other possibilities can be treated with the same formalism. We will describe each state available to the DNA molecule, and ascribe an energy to that state.

We explicitly model only the DNA molecule itself. However, because the energy parameters used as inputs into the model have been determined from *in vitro* experiments, they implicitly include the effects of solvation, ionic conditions, and other environmental factors. The resulting ''effective Hamiltonian'' therefore implicitly incorporates a dependence on these environmental conditions.

More generally, consider a Hamiltonian $H_0(\vec{x},\vec{y})$, where $\vec{x}$ refers to the DNA degrees of freedom which will be explicitly considered, and $\vec{y}$ refers to any other DNA degrees of freedom as well as to the environmental degrees of freedom. The constant temperature and pressure partition function Z of the DNA plus environment [35] is then given by

$$Z = \mathrm{Tr}_{\vec{x},\vec{y}}\, e^{-\beta H_0(\vec{x},\vec{y})} \tag{1}$$

with the Gibbs free energy

$$G = \left(-\frac{1}{\beta}\right)\ln(Z) = \left(-\frac{1}{\beta}\right)\ln[\mathrm{Tr}_{\vec{x},\vec{y}}\, e^{-\beta H_0(\vec{x},\vec{y})}]. \tag{2}$$

Here, $\mathrm{Tr}_{\vec{x},\vec{y}}$ refers to sums or integrals over $\vec{x}$ and $\vec{y}$ as appropriate and $\beta = 1/(k_B T)$, where k_B is Boltzmann's constant and T is the absolute temperature. Equation (1) can also be written in the form

$$Z = \mathrm{Tr}_{\vec{x}}\, e^{-\beta H(\vec{x})}, \tag{3}$$

where

$$H(\vec{x}) = \left(-\frac{1}{\beta}\right) \ln[\mathrm{Tr}_{\vec{y}}\, e^{-\beta H_0(\vec{x},\vec{y})}], \tag{4}$$

giving

$$e^{-\beta H(\vec{x})} = \mathrm{Tr}_{\vec{y}}\, e^{-\beta H_0(\vec{x},\vec{y})}. \tag{5}$$

Because of the form of Eq. (5), $H(\vec{x})$ is sometimes considered to refer to the ''free energy'' of a particular system A plus its environment B, for a fixed configuration of that system A. Alternatively, one can consider $H(\vec{x})$ as an ''effective Hamiltonian,'' with interactions between the $\vec{x}$ degrees of freedom renormalized by the environment (and possibly temperature dependent). We will primarily use the terminology ''effective Hamiltonian'' and ''energy'' in this paper.

For a quantity $\mathcal{O}$ which depends only upon the $\vec{x}$ degrees of freedom $[\mathcal{O}=\mathcal{O}(\vec{x})]$, the expectation value is given by

$$\langle \mathcal{O} \rangle = \frac{\mathrm{Tr}_{\vec{x},\vec{y}}\{\mathcal{O}(\vec{x}) e^{-\beta H_0(\vec{x},\vec{y})}\}}{\mathrm{Tr}_{\vec{x},\vec{y}}\{e^{-\beta H_0(\vec{x},\vec{y})}\}} = \frac{\mathrm{Tr}_{\vec{x}}\{\mathcal{O}(\vec{x})\mathrm{Tr}_{\vec{y}}[e^{-\beta H_0(\vec{x},\vec{y})}]\}}{\mathrm{Tr}_{\vec{x}}\{\mathrm{Tr}_{\vec{y}}[e^{-\beta H_0(\vec{x},\vec{y})}]\}}$$

$$= \frac{\mathrm{Tr}_{\vec{x}}\{\mathcal{O}(\vec{x}) e^{-\beta H(\vec{x})}\}}{\mathrm{Tr}_{\vec{x}}\{e^{-\beta H(\vec{x})}\}}. \tag{6}$$

Thus, as long as an effective Hamiltonian $H(\vec{x})$ is used, expectation values can be calculated as usual. This is the formal basis for the procedure we will follow.

States and their energies

In each state of the DNA molecule the linking difference α is partitioned among three factors. First, the secondary structure is specified by describing which base pairs are denatured in that state. Denaturation decreases the unstressed helicity of the involved base pairs from that characteristic of the B-form duplex to that of the untwisted condition. If there are n denatured base pairs in the molecule in the given state, the total change in unstressed helicity resulting from this transition is $-n/A$, where $A=10.4$ bp/turn for denaturation [36]. Only when $n=-A\alpha$ does the extent of denaturation exactly relax the imposed linking difference. All states for which $n\neq -A\alpha$ will experience some level of uncompensated superhelicity. The resulting torsional stresses cause the two single strands comprising a denatured region to twist around each other. We denote the total change of twist arising from this effect by $\mathcal{T}$. Finally, the residual linking difference α_r is that portion of α that is not expressed by either of the above two structural alterations. According to the formalism presented above, this residual deformation need not be decomposed further, since the energetics associated with α_r have been determined experimentally.

These deformations are all coupled together by the topological constraint arising from the constancy of the linking difference α:

$$\alpha = -\frac{n}{A} + \mathcal{T} + \alpha_r. \tag{7}$$

With α and the secondary structure of each base pair specified, the torsional deformation $\mathcal{T}$ and the residual linking difference α_r still can vary, provided they do so in a reciprocal manner consistent with Eq. (7). We consider in turn each type of deformation and its associated energetics.

Local denaturation

Let n_j, $1\leqslant j\leqslant N$, be a variable whose value is $n_j=1$ when the base pair at position j is denatured (sometimes also called ''separated'' or ''open'') and $n_j=0$ when it is in the B form (i.e., ''bonded'' or ''closed''). Because the molecule is circular, base pairs 1 and N are neighbors. To accommodate this periodic boundary condition, we set $n_{N+j}=n_j$ as needed. Specifying the value of each n_j determines a unique state of secondary structure of the molecule, in which the total number of denatured base pairs is

$$n = \sum_{j=1}^{N} n_j. \tag{8}$$

We denote the energy required to denature base pair j if it is at the edge of an open region by b_j. The values of b_j can be assigned individually to each base pair. In contrast to previous approaches [5], this method places no restrictions on the values they can have. This energy of denaturation is known to vary in complex ways with base sequence and environmental conditions. Values of b_j have been measured experimentally as functions both of base pair composition [6] and of ionic strength [37]. The near-neighbor sequence dependence of the enthalpy and entropy of denaturation have been determined under various environmental conditions [7–9]. Energies of denaturation have been evaluated for methylated bases, and for abasic sites [31,32,38]. Previous theoretical analyses of superhelical DNA denaturation assumed copolymeric transition energetics, with a single value b_{AT} ascribed to each AT base pair, and another value b_{GC} given to each GC pair [5]. Under the environmental conditions of the experiments used to detect superhelical denaturation, these are $b_{AT}=0.26$ kcal/mol and $b_{GC}=1.31$ kcal/mol [16].

A ''run'' is a region composed entirely of separated base pairs. Since n_j only changes with j at the boundary of a run, the number r of runs in a state of a circular molecule can be expressed as

$$r = \sum_{j=1}^{N} n_j(1-n_{j+1}). \tag{9}$$

An initiation energy a is required to nucleate a run of denaturation. This arises in large part from the energy needed to break the extra hydrophobic ''stacking'' interaction that must be disrupted when the first base pair in a run is separated. The initiation energy for denaturation is large, $a \approx 10-13$ kcal/mol, depending on environmental conditions [16,39–42]. In the calculations reported below we use the value $a=10.8$ kcal/mol that is appropriate for the experimental conditions under which superhelical denaturation is measured [16].

Hence the total chemical energy needed to denature the base pairs in the state is

$$H_c = ar + \sum_{j=1}^{N} b_j n_j = \sum_{j=1}^{N} \{(a+b_j)n_j - a n_j n_{j+1}\}. \quad (10)$$

Interstrand twisting within denatured regions

Because single stranded DNA is highly flexible and the denatured regions within a superhelical molecule generally remain torsionally stressed, the unpaired strands comprising them will tend to interwind. If the base pair at site j is denatured ($n_j = 1$), and has a helical twist of τ_j rad/bp, the energy associated with this deformation is

$$H(\tau_j) = \frac{C n_j \tau_j^2}{2}. \quad (11)$$

This energy arises in part from configurational restrictions due to helical interwinding. The value of the torsional stiffness, $C \approx 9.3 \times 10^{-21}$ erg cm, is known from experiments [16,42]. We do not explicitly model fluctuations in the twist of bonded (nondenatured) base pairs, as the torsional stiffness of *B*-form DNA is about two orders of magnitude greater than that of the individual denatured strands [26]. Instead, this effect is subsumed within α_r, the residual superhelicity.

We will consider the torsional deformations τ_j at two levels of detail. In case (1), τ_j is set to τ for each separated base pair j, so that the total twist $\mathcal{T}$ of the open regions is

$$\mathcal{T} = \frac{n\tau}{2\pi}. \quad (12)$$

This is done primarily to enable comparisons with previous treatments [5,16]. In case (2), we allow the τ_j associated with each denatured base pair to fluctuate independently, giving

$$\mathcal{T} = \sum_{j=1}^{N} \frac{n_j \tau_j}{2\pi}. \quad (13)$$

Note that n_j, and hence this summand, is nonzero only at the n denatured base pairs in the state under consideration.

Residual superhelicity and total energy

Once the separated base pairs and their torsional deformations are specified, the residual linking difference is determined as

$$\alpha_r = \alpha + \frac{n}{A} - \mathcal{T}. \quad (14)$$

This residual linking difference is comprised of twisting of the *B*-form regions, as well as bending deformations. However, for present purposes, α_r need not be decomposed into these constituents because the energy associated with it is known from experiments.

The energy associated with superhelical deformations has been measured by several experimental techniques to be quadratic in the linking difference under conditions where no denaturation occurs [43–45]. (In that case $\alpha_r = \alpha$.) The same quadratic functional form also has been experimentally found for the residual linking difference when denaturation does occur [16,42]:

$$H_r = \frac{K\alpha_r^2}{2} = \frac{K}{2}\left(\alpha + \frac{n}{A} - \mathcal{T}\right)^2. \quad (15)$$

The coefficient K has been determined experimentally to vary inversely with molecular length N, having the value $K \approx 2220RT/N$ at the physiological temperature 37 °C.

The total energy associated with a state depends on the manner in which the torsional deformations τ_j are being modeled. Adding all the contributions, we find that the Hamiltonian H_1 for case (1), where $\tau_j = \tau$, is

$$H_1 = \frac{nC\tau^2}{2} + \frac{K}{2}\left(\alpha + \frac{n}{A} - \frac{n\tau}{2\pi}\right)^2 + \sum_{j=1}^{N} \{(a+b_j)n_j - a n_j n_{j+1}\}. \quad (16)$$

In case (2), where each of the n separated base pairs is torsionally deformed at a rate of τ_j rad/bp, the Hamiltonian is

$$H_2 = \sum_{j=1}^{N} \frac{C n_j \tau_j^2}{2} + \frac{K}{2}\left(\alpha + \frac{n}{A} - \sum_{j=1}^{N} \frac{n_j \tau_j}{2\pi}\right)^2 + \sum_{j=1}^{N} \{(a+b_j)n_j - a n_j n_{j+1}\}. \quad (17)$$

Calculation of the partition function

The calculation of the partition function involves summing and integrating the usual Boltzmann factor $e^{-\beta H}$ over all the states available to the system, where H here refers to the Hamiltonian of a given state and $\beta = 1/(k_B T)$. We proceed by first eliminating the degrees of freedom associated with the τ_j's, the twisting of the separated DNA strands.

First, consider case (1), where $\tau_j = \tau$ for each separated base pair. For each number n of separated base pairs and imposed linking difference α, we minimize Eq. (16) with respect to τ. This leads to the condition $\alpha_r K = 2\pi C\tau$. [The same condition follows when Eq. (17) is minimized with respect to the τ_j's.] If we replace τ by this value, as was done in previous work [5,16], the effective Hamiltonian H_1, now dependent only on the n_j's, becomes

$$H_1 = \frac{2\pi^2 CK}{4\pi^2 C + Kn}\left(\alpha + \frac{n}{A}\right)^2 + \sum_{j=1}^{N} \{(a+b_j)n_j - a n_j n_{j+1}\}. \quad (18)$$

Under this minimizing assumption the partition function associated with H_1 has the form

$$Z_1 = \sum_{\mathcal{S}} Q_1(n) e^{-\beta \Sigma_{j=1}^{N} \{(a+b_j)n_j - a n_j n_{j+1}\}}, \quad (19)$$

where

$$Q_1(n)=\exp\left[\frac{-2\pi^2\beta CK}{4\pi^2C+Kn}\left(\alpha+\frac{n}{A}\right)^2\right], \tag{20}$$

and

$$\sum_{\mathcal{S}}=\sum_{n_1=0}^{1}\sum_{n_2=0}^{1}\cdots\sum_{n_N=0}^{1} \tag{21}$$

denotes summation over the 2^N states of secondary structure.

In case (2), where each τ_j corresponding to an open base pair is allowed to fluctuate, the partition function has a similar form, differing only in the prefactor $Q_2(n)$. One can integrate over the n continuous degrees of freedom τ_j to obtain the expression

$$Q_2(n)=\prod_{j=1}^{n}\int_{-\infty}^{\infty}d\tau_j\exp\left[-\beta\left\{\sum_{j=1}^{n}\frac{C\tau_j^2}{2}+\frac{K}{2}\left(\alpha+\frac{n}{A}-\sum_{j=1}^{n}\frac{\tau_j}{2\pi}\right)^2\right\}\right]. \tag{22}$$

Performing a matrix version of the completion of squares, one may evaluate this integral to be

$$Q_2(n)=Q_1(n)\left(\left\{\frac{2\pi}{\beta C}\right\}^n\frac{4\pi^2C}{4\pi^2C+Kn}\right)^{1/2}. \tag{23}$$

In each case the partition function may be expressed as

$$Z_\lambda=\sum_{\mathcal{S}}Q_\lambda(n)\exp\left\{-\beta\sum_{j=1}^{N}[(a+b_j)n_j-an_jn_{j+1}]\right\}, \tag{24}$$

where λ equals 1 or 2 depending on the case being considered. We will henceforth drop the subscript λ unless specifically noted, as our calculation strategy will apply in both cases.

In both cases (1) and (2) the partition function may be written in the form $Z_\lambda=\mathrm{Tr}\{\exp(-\beta\mathcal{H}_\lambda)\}$, where $\mathcal{H}_\lambda$ is a function of the n_j's only. From Eqs. (20) and (23), one of the terms in each $\mathcal{H}_\lambda$ is $\ln\{Q_\lambda(n)\}$, which contains the factor $(\alpha+n/A)^2$. Since $n^2=\Sigma_{j,k=1}^{N}n_jn_k$, this shows that a coupling is induced in $\mathcal{H}_\lambda$ between every pair (j,k) of the base pairs.

A naive calculation of the partition functions of Eq. (24) would require a computation time growing exponentially with the number of sites N. However, an expression having the functional form of Eq. (24) can be evaluated in polynomial time using the following procedure. First we write $Q(n)$ as

$$Q(n)=\sum_{m=0}^{N}\delta_{m,n}Q(m), \tag{25}$$

where $\delta_{m,n}$ is the Kronecker δ function:

$$\delta_{m,n}=\begin{cases}1 & \text{if } m=n\\ 0 & \text{otherwise.}\end{cases} \tag{26}$$

Expressing the δ function in the form

$$\delta_{m,n}=\left(\frac{1}{N+1}\right)\sum_{k=0}^{N}\exp\left(\frac{2\pi ik(n-m)}{N+1}\right), \tag{27}$$

where $i=\sqrt{-1}$, we obtain

$$Q(n)=\left(\frac{1}{N+1}\right)\sum_{m=0}^{N}\sum_{k=0}^{N}Q(m)\exp\left(\frac{2\pi ik(n-m)}{N+1}\right). \tag{28}$$

Placing this expression for $Q(n)$ into Eq. (24), using the fact that $n=\Sigma_{j=1}^{N}n_j$, and rearranging terms, yields

$$Z=\sum_{k=0}^{N}F(k)q(k), \tag{29}$$

where

$$q(k)=\left(\frac{1}{N+1}\right)\sum_{m=0}^{N}Q(m)\exp\left(-\frac{2\pi ikm}{N+1}\right) \tag{30}$$

and

$$F(k)=\sum_{\mathcal{S}}\exp\{-\beta\mathcal{H}(k)\}, \tag{31}$$

with

$$\mathcal{H}(k)=\sum_{j=1}^{N}c_j(k)n_j-an_jn_{j+1} \tag{32}$$

and

$$c_j(k)=a+b_j-\frac{2\pi ik}{\beta(N+1)}. \tag{33}$$

$\mathcal{H}(k)$ has the form of a one-dimensional lattice gas (or, equivalently, Ising model), in which the chemical potential (magnetic field) is site dependent and complex.

The $F(k)$ term derives only from the chemical energies associated with base pair separations, while the $q(k)$ term depends upon mechanical parameters associated with topological and geometric factors. These are the imposed linking difference α, the residual superhelicity α_r, and the torsional deformations of the denatured regions. Separating the factors arising from the denaturation transition from those derived from the topological constraint enables efficient evaluation of the entire expression. $F(k)$ requires a summation over all states $\mathcal{S}$ of the secondary structure of the molecule. This summation will be performed using the transfer matrix

method [46], as described below. Once $F(k)$ and the appropriate prefactor $Q(m)$ have been evaluated, the balance of the computation is straightforward. We will show that all the expectation values of interest can be calculated using equations having the general form of Eqs. (29)–(33).

Transfer matrix method

We begin by briefly reviewing the transfer matrix method [46], originally formulated for the Ising model. First, suppose $\mathbf{M}_1,\ldots,\mathbf{M}_N$ are $q\times q$ square matrices. We number rows and columns of these matrices from 0 to $q-1$, and here denote the element in the ith row and the jth column of $\mathbf{M}_l$ by $m^l_{i,j}$. The product of all these matrices is

$$\mathbf{P}=\mathbf{M}_1\mathbf{M}_2\cdots\mathbf{M}_N=\prod_{l=1}^{N}\mathbf{M}_l, \tag{34}$$

and the trace of $\mathbf{P}$ is

$$\mathrm{Tr}(\mathbf{P})=\sum_{j=0}^{q-1}p_{jj}=\sum_{j_0=0}^{q-1}\cdots\sum_{j_{N-1}=0}^{q-1}(m^1_{j_0,j_1}\cdots m^l_{j_{l-1},j_l}\cdots m^N_{j_{N-1},j_0}). \tag{35}$$

To illustrate the transfer matrix method, we calculate the $F(k)$ of Eq. (31) needed to compute the partition function. We set

$$\mathcal{H}(k)=\sum_{l=1}^{N}\mathcal{H}_l(k), \tag{36}$$

where we here choose $\mathcal{H}_l(k)$ to have the symmetric form

$$\mathcal{H}_l(k)=\tfrac{1}{2}\{c_l(k)n_l+c_{l+1}(k)n_{l+1}\}-an_ln_{l+1}, \tag{37}$$

which is the form we will use in our algorithmic implementation. (This choice is somewhat arbitrary; we know of no particular advantage to using a symmetric versus a non-symmetric form.) Each $\mathcal{H}_l(k)$ only depends upon the variables n_l and n_{l+1}. Because the collection of all states of the system $\mathcal{S}$ is exhausted by permitting each variable n_l, $l=1,...,N$, to take on every possible value, it follows that

$$F(k)=\sum_{\mathcal{S}}e^{-\beta\mathcal{H}(k)}=\sum_{n_1=0}^{1}\cdots\sum_{n_N=0}^{1}(e^{-\beta\mathcal{H}_1(k)[n_1,n_2]}\cdots e^{-\beta\mathcal{H}_l(k)[n_l,n_{l+1}]}\cdots e^{-\beta\mathcal{H}_N(k)[n_N,n_1]}). \tag{38}$$

In this form $F(k)$ has the same structure as that given in Eq. (35), with $q=2$. This shows that $F(k)$ can be expressed as the trace of the product of 2×2 transfer matrices $\mathbf{M}_l(k)$, $l=1,\ldots,N$, one for each base pair in the molecule. The transfer matrix $\mathbf{M}_l(k)$ has entry $m(k)^l_{i,j}$ ($i=0,1$ and $j=0,1$) corresponding, respectively, to the values of $e^{-\beta\mathcal{H}_l(k)}$ arising when $n_l=0,1$ and $n_{l+1}=0,1$. Thus, the matrix $\mathbf{M}_l(k)$ that occurs when one uses the symmetric form of $\mathcal{H}_l(k)$ in the evaluation of $F(k)$ is

$$\mathbf{M}_l(k)=\begin{pmatrix}1 & e^{-\beta c_{l+1}(k)/2}\\ e^{-\beta c_l(k)/2} & e^{-\beta(-a+\{c_l(k)+c_{l+1}(k)\}/2)}\end{pmatrix}. \tag{39}$$

This shows that the function $F(k)$ in Eq. (31) may be expressed as

$$F(k)=\mathrm{Tr}\left(\prod_{l=1}^{N}\mathbf{M}_l(k)\right). \tag{40}$$

Computations using this method require an evaluation of products of large numbers of matrices. In the standard one-dimensional Ising model, the transition energetics are identical at every position, so the transfer matrices are all the same and the trace can be expressed as the sum of powers of eigenvalues [46]. In the present case the energy of transition varies with base sequence, so the transfer matrices $\mathbf{M}_l(k)$ associated with different base pairs will not be identical. As multiplication of these matrices does not commute, this operation must be performed numerically. The numerical implementation of this approach is described in a later section.

Calculation of ensemble averages

The ensemble average values of several quantities provide important insights into the transition behavior of a superhelical DNA molecule. These include the average numbers $\overline{n}$ of denatured base pairs and $\overline{r}$ of runs of transition, and the average total twist $\overline{\mathcal{T}}$ of the denatured regions. The ensemble average residual linking difference $\overline{\alpha_r}$, obtainable from $\overline{\mathcal{T}}$, is an important parameter to calculate because it can be experimentally measured using gel electrophoresis techniques [42]. The biologically most interesting information involves the locations where denaturation occurs, and the relative extents and energy costs of transition at those locations. We evaluate locations and extents of transition by calculating the probability of transition $p_l=\overline{n_l}$ individually for each base pair $1\leqslant l\leqslant N$. The resulting transition profile is typically displayed graphically by plotting p_l against sequence location l. (An example is displayed as the upper graph in Fig. 4 below.) A method to calculate the destabilization energies of individual base pairs is described in a later section.

The additional quantities that must be calculated in order to evaluate these ensemble averages all have the functional forms expressed in Eqs. (29)–(33). Only the forms of the prefactors $Q(m)$ and the summands of $F(k)$ may differ. Hence they can be evaluated by the general procedure that was described above for the partition function.

Calculation of the average number $\overline{n}$ of separated base pairs

The ensemble average number of separated base pairs $\overline{n}$ is

$$\overline{n}=\frac{Z(n)}{Z}, \tag{41}$$

where

$$Z(n)=\sum_{\mathcal{S}} nQ(n)e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}. \tag{42}$$

This expression may be evaluated using the same technique as was described above for the partition function, with the sole modification that $Q(n)$ is replaced by $nQ(n)$. As this is still a function of n alone, the procedure applies unchanged. Alternatively, one may evaluate $\overline{n}$ as the sum of the transition probabilities $\overline{n_l}$ of the individual base pairs, which we now consider.

Calculation of $\overline{n_l}$

The transition profile of a DNA sequence at linking difference α is a graph of the equilibrium probability of transition p_l, $1\leqslant l\leqslant N$, of each base pair in the molecule. Here $p_l=\overline{n_l}$ is given by

$$\overline{n_l}=\frac{Z(n_l)}{Z}, \tag{43}$$

where $Z(n_l)$ is the contribution to the partition function from all states in which base pair l is separated:

$$Z(n_l)=\sum_{\mathcal{S}} Q(n)n_le^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}. \tag{44}$$

$Z(n_l)$ may be cast in the functional form of Eq. (29), with $F(k)$ replaced by

$$F_l(k)=\sum_{\mathcal{S}} n_l\exp\left(-\beta\sum_{j=1}^{N}\mathcal{H}_j(k)\right). \tag{45}$$

Here the $\mathcal{H}_j(k)$'s are the same as those used in the evaluation of the partition function. This has the effect of replacing the transfer matrix $\mathbf{M}_l(k)$ by that corresponding to $n_l\exp(-\beta\mathcal{H}_l(k))$,

$$\mathbf{M}_l'(k)=\begin{pmatrix}0 & 0\\ e^{-\beta c_l(k)/2} & e^{-\beta(-a+\{c_l(k)+c_{l+1}(k)\}/2)}\end{pmatrix}. \tag{46}$$

All the other $N-1$ transfer matrices remain unchanged, as do the $q(k)$'s. Because $Z(n_l)$ is the partition function with base pair l always separated, one can calculate the free energy $\Delta G(l)$ required to separate base pair l as

$$\Delta G(l)=(-k_BT)\{\ln Z(n_l)-\ln Z\}=-\frac{\ln p_l}{\beta}. \tag{47}$$

Calculation of the average number $\overline{r}$ of runs

The ensemble average number of runs of transition, $\overline{r}$, is given by

$$\overline{r}=\frac{Z(r)}{Z}, \tag{48}$$

where

$$Z(r)=\sum_{\mathcal{S}} Q(n)re^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}. \tag{49}$$

Expressing r as

$$r=\sum_{l=1}^{N}(n_l-n_ln_{l+1}), \tag{50}$$

one obtains

$$Z(r)=\sum_{l=1}^{N}[Z(n_l)-Z(n_ln_{l+1})], \tag{51}$$

where

$$Z(n_ln_{l+1})=\sum_{\mathcal{S}} Q(n)n_ln_{l+1}e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}. \tag{52}$$

The expressions $Z(n_l)$ have been calculated above for each l. The term $Z(n_ln_{l+1})$ again has the functional form of Eq. (29), with $F(k)$ now replaced by

$$F_{l,l+1}(k)=\sum_{\mathcal{S}} n_ln_{l+1}\exp\left(-\beta\sum_{j=1}^{N}\mathcal{H}_j(k)\right). \tag{53}$$

In this case the transfer matrix $\mathbf{M}_l(k)$ is replaced by that corresponding to $n_ln_{l+1}\exp(-\beta\mathcal{H}_l(k))$,

$$\mathbf{M}_l''(k)=\begin{pmatrix}0 & 0\\ 0 & e^{-\beta(-a+\{c_{l+1}(k)+c_l(k)\}/2)}\end{pmatrix}, \tag{54}$$

while all other transfer matrices remain unchanged. [Equivalently, one could replace $\exp(-\beta\mathcal{H}_l(k))$ by $n_l\exp(-\beta\mathcal{H}_l(k))$ and $\exp(-\beta\mathcal{H}_{l+1}(k))$ by $n_{l+1}\exp(-\beta\mathcal{H}_{l+1}(k))$, giving two matrices having the form of Eq. (46).]

Calculation of the average total twist $\overline{\mathcal{T}}$

We now consider the ensemble average total twist $\overline{\mathcal{T}}$ of the unpaired (open) regions, which we write in the form

$$\overline{\mathcal{T}}=\frac{1}{2\pi}\sum_{j=1}^{N}\overline{n_j\tau_j}=\frac{Z(\mathcal{T})}{Z}. \tag{55}$$

For both cases (1) and (2) ($\lambda=1$ and 2), we find that

$$Z_\lambda(\mathcal{T})=\sum_{\mathcal{S}} Q_{\lambda,\tau}(n)e^{-\beta\Sigma_j\{(a+b_j)n_j-an_jn_{j+1}\}}, \tag{56}$$

where

$$Q_{\lambda,\mathcal{T}}(n)=Q_\lambda(n)\left(\frac{Kn}{4\pi^2C+Kn}\right)\left(\alpha+\frac{n}{A}\right). \tag{57}$$

The expressions $Q_{\lambda,\mathcal{T}}$ that result are functions of n alone, so the evaluations of $Z(\mathcal{T})$ proceed in the same manner as was described above for the partition function.

Calculation of residual twisting $\overline{\alpha_r}$

Using Eq. (7) to express α_r in terms of the other deformations, one obtains

$$\overline{\alpha_r}=\alpha+\frac{\overline{n}}{A}-\overline{\mathcal{T}}. \tag{58}$$

The procedures used to calculate the two average values on the right hand side of this equation have already been described.

Calculation of "energy" $\overline{H}$

We now calculate the $\overline{H_\lambda}$'s, the ensemble averages of the effective Hamiltonians of Eqs. (16) and (17). These averages are used in the calculations of base pair destabilization energies, discussed below. As a reminder, $\lambda=1$ and 2 refer to our different assumptions regarding twisting deformations, with $\lambda=2$ the more generally physically relevant.

For case (1) ($\lambda=1$), we use the H_1 of Eq. (16) to obtain

$$\overline{H}=\frac{\Sigma_{\mathcal{S}}H_1e^{(-\beta H_1)}}{\Sigma_{\mathcal{S}}e^{(-\beta H_1)}}=\frac{Z_1(H)}{Z_1}. \tag{59}$$

The partition function Z_1 has already been calculated. $Z_1(H)$ is given by

$$Z_1(H)=\sum_{\mathcal{S}}\{\mathcal{R}_1(n)e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}\}+\sum_{j=1}^{N}(a+b_j)Z_1(n_j)-\sum_{j=1}^{N}aZ_1(n_jn_{j+1}), \tag{60}$$

where, from Eq. (18),

$$\mathcal{R}_1(n)=\left[\frac{2\pi^2CK}{4\pi^2C+Kn}\left(\alpha+\frac{n}{A}\right)^2\right]Q_1(n) \tag{61}$$

As the first term of Eq. (60) has the general form of the partition function of Eq. (24), and $Z_1(n_j)$ and $Z_1(n_jn_{j+1})$ have been evaluated above, $\overline{H_1}$ can be calculated using the procedure developed above for Eqs. (29)–(33).

Integrating over the τ_j's, we find that case (2) reduces to the form of Eqs. (60) and (61) as well. Specifically,

$$\overline{H}=\frac{Z_2(H)}{Z_2}. \tag{62}$$

Z_2 has already been evaluated, and $Z_2(H)$ is given by

$$Z_2(H)=\sum_{\mathcal{S}}\{\mathcal{R}_2(n)e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_ln_{l+1}\}}\}+\sum_{j=1}^{N}(a+b_j)Z_2(n_j)-\sum_{j=1}^{N}aZ_2(n_jn_{j+1}), \tag{63}$$

where

$$\mathcal{R}_2(n)=Q_2(n)\left[\frac{n}{2\beta}+\frac{2\pi^2CK}{4\pi^2C+Kn}\left(\alpha+\frac{n}{A}\right)^2\right]. \tag{64}$$

Calculation with fixed base pair separations

Several types of externally imposed conditions may affect the secondary structure of specific base pairs *in vivo*. Site-specific DNA binding proteins or enzymes may hold particular base pair(s) either open ($n_l=1$) or closed ($n_l=0$). Alternatively, abasic sites are created when the purine or pyrimidine base at a site is lost. This does not disrupt the continuity of the sugar-phosphate backbone, so the topological constraint is unaffected. However, there being no base at that site, Watson-Crick pairing is impossible. Enforced openings or closures can have a significant effect on the destabilization experienced by other base pairs throughout the domain. For example, the externally enforced separation of a base pair, as occurs at an abasic site, permanently nucleates denaturation at that site, so the large initiation energy a needed to start a run at any other position is not needed there. This increases the probability that additional denaturation will occur in that region over what would be expected in the intact molecule [23].

The partition function that arises when base pair l is constrained to be open is given by

$$Z(n_l)=\sum_{\mathcal{S}} Q(n)n_le^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}} \tag{65}$$

which is a sum only over states where $n_l=1$. Similarly, the partition function with base pair l held closed is

$$\sum_{\mathcal{S}} Q(n)(1-n_l)e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}=Z-Z(n_l). \tag{66}$$

If base pairs l and l' are both held open, for example, the partition function becomes

$$Z(n_l,n_{l'})=\sum_{\mathcal{S}} Q(n)n_ln_{l'}e^{-\beta\Sigma_{j=1}^{N}\{(a+b_j)n_j-an_jn_{j+1}\}}. \tag{67}$$

The ensemble averages derived previously can all be calculated with these base pairing constraints. For example, the probability $\overline{n_l}(l')$ that site l is separated when site l' is held open ($n_{l'}$ fixed at 1) is given by

$$\overline{n_l}(l') = \frac{Z(n_l, n_{l'})}{Z(n_{l'})}. \tag{68}$$

The probability that site l is separated with the base pair at site l' held closed is similarly given by

$$\frac{Z(n_l) - Z(n_l, n_{l'})}{Z - Z(n_{l'})}. \tag{69}$$

For certain averages and under general base pair separation constraints it is necessary to calculate quantities of the general form

$$Z(n_{l_1}, n_{l_2} \ldots n_{l_M}) = \sum_{\mathcal{S}} Q(n) n_{l_1} n_{l_2} \ldots n_{l_M} e^{-\beta \Sigma_{j=1}^{N} \{(a+b_j) n_j - a n_j n_{j+1}\}}. \tag{70}$$

This is accomplished by modifying the appropriate transfer matrices, as has been done in Eqs. (46) and (54) above.

Calculation of destabilization energies

Sites where stress-induced destabilization occurs but is not sufficient to drive denaturation may be important as targets for other molecules, such as helicases, whose activities affect strand separation. If these other molecules can contribute only marginally to the energy needed to denature their target DNA site, they may be able to induce separation only when that site is already partially destabilized. In this way imposed torsional stresses (such as those produced by negative superhelicity) may facilitate events involving denaturation, even where these stresses do not drive separation to completion. The calculations of destabilization energies developed here are designed to find such partially destabilized regions.

One estimator of the energy required to separate a particular base pair l can be obtained by comparing the usual ensemble average $\overline{H}$ with the ensemble average $\overline{H}(l)$ found when base pair n_l is held open. The difference $\Delta\overline{H}(l) = \overline{H}(l) - \overline{H}$ provides a measure of the extent to which base pair l is destabilized by the imposed stresses: the smaller this difference, the more destabilized the base pair [28].

This effective destabilization energy $\overline{H}(l)$ must be calculated for each site l. Restoring the λ subscript referring to twisting assumptions, we have

$$\overline{H}_{\lambda}(l) = \frac{Z(n_l, H_{\lambda})}{Z(n_l)}, \tag{71}$$

with

$$Z(n_l, H_{\lambda}) = \sum_{\mathcal{S}} \{\mathcal{R}_{\lambda}(n) n_l e^{-\beta \Sigma_{j=1}^{N} \{(a+b_j) n_j - a n_j n_{j+1}\}}\} + \sum_{j=1}^{N} (a+b_j) Z_{\lambda}(n_j n_l) - \sum_{j=1}^{N} a Z_{\lambda}(n_j n_{j+1} n_l). \tag{72}$$

The methods to evaluate each term in this expression have been described above. The $\Delta\overline{H}(l)$'s are typically less noisy estimators of site-specific destabilization energies than the $\Delta G(l)$'s of Eq. (47) [28]; however, they are also computationally more expensive.

Evaluation of an alternate strategy for treating the linking number constraint

As seen from Eq. (7), we impose an exact constraint on the linking number α. Another possible approach could be to use a ''linking number potential'' (LNP), which we denote by μ. This strategy would be implemented by retaining α_r as an independent variable, using $K\alpha_r^2/2$ in Eqs. (16) and (17) instead of the rightmost term of Eq. (15), adding a term $-\mu[-(n/A) + \mathcal{T} + \alpha_r]$ to the effective Hamiltonians of Eqs. (16) and (17), and then adjusting μ until the expectation value of the right side of Eq. (7) achieves the desired value of α. This type of approach, analogous to going for example from the canonical to the grand canonical ensemble, can be effective in cases where the Hamiltonians are homogeneous (at least on some length scale) in the limit of large systems. One example of such a ''thermodynamic limit'' system is a homogeneous Ising ring of a few thousand sites, where a constraint on total spin and the use of the appropriate magnetic ''potential'' would give practically indistinguishable results. However, this approach is not well controlled for the systems we treat.

First, this approach is not strictly applicable in our case because the linking number is not an extensive variable in the usual sense. That is, there is no intensive ''linking density'' whose integration over the molecule yields L. To see this, consider two curves. The first is a figure eight with a single contact point, and the second is the same curve after a strand passage through the contact point. As these conformations differ only infinitesimally, all intensive parameters describing them also differ infinitesimally. So the integrals of any intensive quantities, as they are taken over a fixed and finite length, will also differ infinitesimally. But the linking numbers of these configurations differ by 2. It follows that the linking number is not generally computable from an intensive density. Instead it is expressed using a Gaussian double integral. This means it does not depend on strictly local quantities, but rather on how each part of the molecule is positioned relative to every other part [47].

Second, the thermodynamic limit itself is much more problematic in our model, in part because nonrandom heterogeneity can lead to situations where only a small part or parts of the system are active. As a simple example, inserting 100 *AT* base pairs into a superhelical circular DNA molecule

consisting of 5000 *GC* base pairs would have a dramatic effect on whether and where denaturation would occur, the property we are interested in. But the addition of 100 sites to a 5 K bp homogeneous Ising ring would in general have a negligible physical effect. This behavior has been observed experimentally: in a circular 4000 base pair plasmid, the removal of a particular 16 consecutive base pairs was shown to dramatically affect when and where local denaturation occurred [15]. In general, it is not clear that the systems we study are in the ''thermodynamic limit'' in the usual sense.

Third, the use of a linking number potential μ requires taking weighted averages over ranges of linking differences α. However, α can only assume values that differ by discrete integers, while use of a LNP μ places no such constraints on the right side of Eq. (7). Moreover, when local denaturation first occurs, the case in which we are often most interested, experimental systems (and our model) are very sensitive to the precise value of the linking number, with small (integer-valued) changes having large effects on the denaturation behavior. Under these circumstances it is not clear how accurate a weighted average over a continuous range of linking differences would be for our (finite) systems.

Lastly and perhaps most importantly, as discussed following Eq. (24), direct long-range interactions between all the base pairs are generated in our model when the twist variables are integrated over. However, only nearest-neighbor interactions between base pairs are generated if one uses the LNP approach and integrates over α_r and the twist variables. In general, it is not a well controlled strategy to try to calculate the effects of long-range interactions using a model that contains only nearest-neighbor interactions.

Using a ''linking number potential'' approach might lead to accurate results some of the time, perhaps often. However, we do not in general know how well controlled such an approach is for our systems. Indeed, it seems most likely to fail precisely in the cases in which we have the greatest interest. In fact, the only way to reliably test an LNP approach would be to compare with a numerically exact method such as the one presented in this paper. Further, we develop below a fast, approximate but well-controlled and very accurate implementation of the preceding exact method that can simulate problems of biological interest (molecules several kilobase pairs long) in a few hours on a high-end work station. This makes the implementation of a fast but uncontrolled approximate approach even less important.

ALGORITHM IMPLEMENTATION

Operations count

In this analysis all quantities requiring calculation are sums having the general structure of Eq. (29):

$$\sum_{k=0}^{N} \mathcal{F}(k)\rho(k). \tag{73}$$

Although the $\rho(k)$'s generally will differ in the calculation of different quantities, all involve a discrete Fourier transform of $N+1$ terms [see, e.g., Eq. (30)]. The $\mathcal{F}(k)$'s, which will also typically vary in calculations of different quantities, are expressed as traces of products of N 2×2 transfer matrices.

To describe how these calculations are implemented numerically, we consider three illustrative cases—the calculation of the ensemble average total twist $\overline{\mathcal{T}}$, of the transition profile (which involves evaluating all the $\overline{n_l}$'s, $1\leqslant l\leqslant N$), and of the average number of runs $\overline{r}=\Sigma_{j=1}^{N}(\overline{n_j}-\overline{n_j n_{j+1}})$. The techniques needed in these cases also apply to all others. We will show how all of these calculations together can be performed in $O(N^2)$ steps with $O(N)$ memory. [A minimal memory of $O(N)$ is required simply for storing the base pair sequence.]

We consider first the calculation of the partition function Z and the total twist $\overline{\mathcal{T}}$. In the calculation of Z, the $\rho(k)$ of Eq. (73) is given by the $q(k)$ of Eq. (30). In calculating $\overline{\mathcal{T}}=Z(\mathcal{T})/Z$, the $\rho(k)$ for $Z(\mathcal{T})$ derives from the Fourier transform of the $Q_{\mathcal{T}}(n)$ of Eq. (57). For both Z and $Z(\mathcal{T})$, $\mathcal{F}(k)$ is given by the $F(k)$ in Eqs. (31)–(33).

Each $\mathcal{F}(k)$ in the sum of Eq. (73) requires $O(N)$ operations to evaluate, as it involves the multiplication of N 2×2 matrices. Calculating each $\rho(k)$ also requires no more than $O(N)$ operations. As there are $N+1$ different values of k, a total of $O(N^2)$ operations is required to compute Z and $\overline{\mathcal{T}}$. [The prefactor can be reduced somewhat by using the fast Fourier transform, which requires roughly $O(N\ln N)$ rather than $O(N^2)$ operations to compute all $N+1$ $\rho(k)$'s; however, $O(N^2)$ total operations are still required for the $\mathcal{F}(k)$'s.]

We now consider the transition profile and the total number of runs $\overline{r}$. The N different $Z(n_l)$'s and $Z(n_l n_{l+1})$'s, necessary to compute the $\overline{n_l}$'s and $\overline{r}$, are given by

$$Z(n_l)=\sum_{k=0}^{N} F_l(k)q(k) \tag{74}$$

and

$$Z(n_l n_{l+1})=\sum_{k=0}^{N} F_{l,l+1}(k)q(k), \tag{75}$$

with $F_l(k)$ and $F_{l,l+1}(k)$ given by Eqs. (45) and (53), respectively. A direct calculation of these $2N$ quantities would involve $O(N^3)$ operations. However, this can be reduced to $O(N^2)$ by using the fact that the transfer matrices whose products must be evaluated have a high degree of similarity. Specifically, the matrix products used to evaluate $F(k)$, $F_l(k)$, and $F_{l,l+1}(k)$ differ only in the matrix at position l, as was noted previously [see Eqs. (46) and (54)].

To take advantage of this similarity, we compute two sets of matrix products $\mathbf{P}_l^{(L)}$ and $\mathbf{P}_l^{(R)}$, $1\leqslant l\leqslant N$. $\mathbf{P}_l^{(L)}$ is the ''left'' product of the transfer matrices from 1 to l,

$$\mathbf{P}_l^{(L)}=\mathbf{M}_1\cdots\mathbf{M}_l=\mathbf{P}_{l-1}^{(L)}\cdot\mathbf{M}_l. \tag{76}$$

$\mathbf{P}_l^{(R)}$ is the ''right'' product of the transfer matrices from l to N,

$$\mathbf{P}_l^{(R)}=\mathbf{M}_l\cdots\mathbf{M}_N=\mathbf{M}_l\cdot\mathbf{P}_{l+1}^{(R)}. \tag{77}$$

Recursive evaluation of all the $\mathbf{P}_l^{(L)}$ and $\mathbf{P}_l^{(R)}$ matrices involves $O(N)$ operations, and their storage requires $O(N)$ space. Once these matrices have been calculated, $F_l(k)$ may then be evaluated as

$$F_l(k)=\mathrm{Tr}(\mathbf{P}_{l-1}^{(L)}\cdot\mathbf{M}_l'\cdot\mathbf{P}_{l+1}^{(R)}), \tag{78}$$

and $F_{l,l+1}(k)$ as

$$F_{l,l+1}(k)=\mathrm{Tr}(\mathbf{P}_{l-1}^{(L)}\cdot\mathbf{M}_l''\cdot\mathbf{P}_{l+1}^{(R)}), \tag{79}$$

with $\mathbf{M}_l'$ and $\mathbf{M}_l''$ given by Eqs. (46) and (54). This again requires $O(N)$ operations for all N values of l. Hence, using this approach, all N of the $F_l(k)$'s and $F_{l,l+1}(k)$'s for a given l may be computed in $O(N)$ time using $O(N)$ memory. Therefore, the separation probabilities $\overline{n_l}$ and average number of runs $\overline{r}$ can all be calculated in $O(N^2)$ time with $O(N)$ memory. This matrix multiplication procedure is numerically stable.

A possible alternative procedure with the same time and memory scaling would be to define

$$\mathcal{P}_l=\mathbf{M}_l\cdots\mathbf{M}_N\cdot\mathbf{M}_1\cdots\mathbf{M}_{l-1}. \tag{80}$$

Starting with

$$\mathcal{P}_1=\mathbf{M}_1\cdots\mathbf{M}_N, \tag{81}$$

one could recursively calculate all the $\mathcal{P}_l$'s:

$$\mathcal{P}_l=(\mathbf{M}_{l-1})^{-1}\cdot\mathcal{P}_{l-1}\cdot\mathbf{M}_{l-1}. \tag{82}$$

Using the cyclic property of the trace, one then has

$$F_l(k)=\mathrm{Tr}(\mathbf{M}_l'\cdot(\mathbf{M}_l)^{-1}\cdot\mathcal{P}_l) \tag{83}$$

and

$$F_{l,l+1}(k)=\mathrm{Tr}(\mathbf{M}_l''\cdot(\mathbf{M}_l)^{-1}\cdot\mathcal{P}_l). \tag{84}$$

However, because of the repeated application of *both* $\mathbf{M}_l$'s and $(\mathbf{M}_l)^{-1}$'s, this alternative procedure is numerically unstable.

It requires an additional $O(N^2)$ steps to compute the above quantities for a given set of imposed base pair separations or closures. Hence it requires $O(N^3)$ steps and $O(N^2)$ memory to calculate all N destabilization energies $\Delta\overline{H}(l)$, $1\leqslant l\leqslant N$, from Eqs. (71) and (72).

One can use the following procedure to reduce the number of operations required by approximately a factor of 2. We will illustrate with the calculation of the partition function $Z=\Sigma_{k=0}^{N}F(k)q(k)$ since, as discussed above, calculations of all the observables involve similar techniques.

Because m is an integer, it follows from Eq. (30) that

$$q(N+1-k)=q^*(k), \tag{85}$$

where the star denotes complex conjugation. Similarly,

$$F(N+1-k)=F^*(k), \tag{86}$$

since $n=\Sigma_{j=1}^{N}n_j$ can also assume only integer values. If N is even, we hence have

$$\sum_{k=0}^{N}F(k)q(k)=F(0)q(0)+\sum_{k=1}^{N/2}F(k)q(k)+\sum_{k=1}^{N/2}F^*(k)q^*(k)=F(0)q(0)+2\sum_{k=1}^{N/2}\mathrm{Re}\{F(k)q(k)\}, \tag{87}$$

where $\mathrm{Re}(z)$ denotes the real part of a complex number z. If N is odd, we obtain

$$\sum_{k=0}^{N}F(k)q(k)=F(0)q(0)+F\left(\frac{N+1}{2}\right)q\left(\frac{N+1}{2}\right)+2\sum_{k=1}^{(N-1)/2}\mathrm{Re}\{F(k)q(k)\}. \tag{88}$$

Since only half the k values are now required, the computational time is halved.

As mentioned above, the $\mathcal{F}(k)$ terms in Eq. (73) derive from the chemical properties of denaturation, while the $\rho(k)$'s are determined by mechanical properties associated with the twisting deformations and residual superhelicity. Thus the same set of $\mathcal{F}(k)$'s can be used with different $\rho(k)$'s if only twisting parameters are changed (and vice versa). In particular, once the $\mathcal{F}(k)$'s have been calculated for one value of the linking difference, they do not need to be recalculated to handle additional values. So the incremental cost of treating a range of linking differences is reduced.

Catastrophic cancellation

This algorithm can experience a severe sign cancellation problem when applied to large DNA molecules. A loss of precision occurs when certain summations are performed, because the magnitude of the final sum is much smaller than the magnitude of the largest term in its summand. If the ratio of the total sum to this largest term becomes smaller in magnitude than machine precision, then the final sum will consist only of round-off noise. In our calculations this ratio becomes small exponentially with molecular length.

To illustrate how this problem arises, consider a simple example. Suppose that the energy of initiation a is zero and that the base pair separation energies b_j are also zero. Further, assume that $Q(n)$ has the form

$$Q(n)=e^{-\kappa n}, \tag{89}$$

with $\kappa>0$. This simplified $Q(n)$ shares the essential feature with the $Q(n)$'s of Eqs. (20) and (23) in that it decays with an exponent that is asymptotically linear in n, for large n. With these assumptions, the partition function Z becomes

$$Z=\sum_{S}Q(n)=\sum_{S}e^{-\kappa\Sigma_{j=1}^{N}n_j}=(1+e^{-\kappa})^N. \tag{90}$$

In our algorithm Z is written in the form of Eq. (29),

$$Z=\sum_{k=0}^{N} z(k), \tag{91}$$

where

$$z(k)=F(k)q(k). \tag{92}$$

Under the conditions of this example, $z(0)$ may be shown to have the real part of greatest magnitude of all the $z(k)$'s. From Eqs. (31)–(33), one finds that

$$F(0)=2^N. \tag{93}$$

Also,

$$(N+1)q(0)=\sum_{m=0}^{N} e^{-\kappa m}=\frac{1-e^{-\kappa(N+1)}}{1-e^{-\kappa}}>1, \tag{94}$$

so that

$$z(0)=F(0)q(0)>\left(\frac{1}{N+1}\right)2^N. \tag{95}$$

Hence, the ratio of the partition function Z to its largest term $z(0)$ obeys

$$\frac{Z}{z(0)}<(N+1)\left(\frac{1+e^{-\kappa}}{2}\right)^N, \tag{96}$$

which decays exponentially with system size. For any fixed machine precision, there will be a molecular length N beyond which the calculation of Z will consist entirely of round-off error.

Sample calculations have been performed to determine the extent of this problem in realistic cases. Molecules of varying lengths were analyzed, each at a linking difference $\alpha=-0.055L_0$ which corresponds to physiological levels of superhelicity in bacteria. The energy parameters used in these calculations are the ones which have been shown to apply in the environmental conditions of the experimental procedure of Kowalski and co-workers, by which superhelical denaturation is detected [16,48]. The sequences analyzed were the first N base pairs (i.e., $j=1\ldots,,N$) of the pBR322 DNA molecule, for $100\leqslant N\leqslant 2400$. The transition probability p_l of each base pair was calculated, both in quadratic precision (on a 32 bit machine) and using the arbitrary precision FORTRAN package developed by Bailey [49] with 200 decimal digits of precision. The degradation of precision of the quadratic precision calculation with molecular length was measured by determining the number of decimal digits of agreement between the p_l's calculated each of the two ways, and selecting its minimum value over the sequence analyzed $1\leqslant l\leqslant N$. (We note that the measured average number of decimal digits of agreement over the entire sequence differed from this minimum value by less than 1% in all calculations.)

The results of this procedure are shown in Fig. 1. The number of digits of accuracy of the quadratic precision calculation fell from the maximum of 33.7 that is available in the Hewlett-Packard implementation of quadratic precision to zero at a molecular length of $N=2400$ base pairs. The observed decrease in the number of significant digits was nearly linear with molecular length, as the regression line shows, in qualitative agreement with the exponential form of Eq. (96).

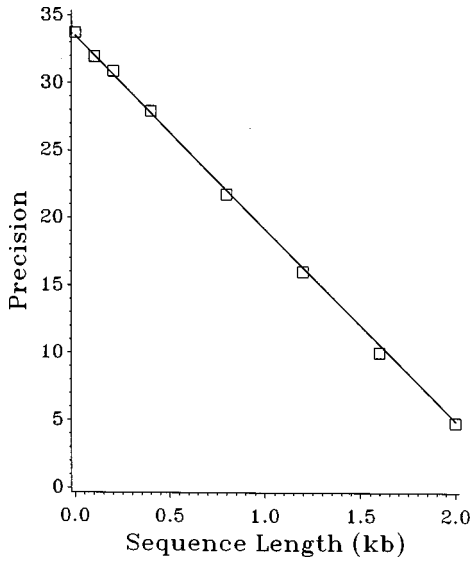

FIG. 1. Stress-induced denaturation is analyzed in molecules of different lengths whose linking differences are chosen so that each has a superhelix density $\sigma=\alpha/L_0=-0.055$, a common physiological value. This analysis is performed using the exact method in both quadratic precision and in high precision with 200 decimal digits of accuracy. The number of significant digits of accuracy of the quadratic precision implementation was assessed by comparing its extent of agreement with the high precision results in each case. The figure plots this accuracy at various sequence lengths as squares, and the fitted regression line is also shown. One sees a rapid, linear, loss in the number of significant digits with sequence length due to catastrophic cancellation. Extrapolation of this regression line allows one to estimate the arithmetic precision needed to achieve a prescribed accuracy in the analysis of a DNA domain of any length.

These results demonstrate the need to implement this algorithm in high precision arithmetic. The calculations needed to analyze a 10 000-bp molecule under these conditions can be estimated from these sample runs to suffer the loss of approximately 140 decimal digits of accuracy due to catastrophic cancellation. Thus high precision implementations using floating point arithmetic with 200 decimal digits of accuracy will generally suffice to analyze biological sequences of this size. Our calculations were implemented using Bailey's multiprecision FORTRAN package MPFUN, which allows the user to specify the level of precision [49].

Implementing these calculations using multiprecision arithmetic significantly slows their execution speed. The CPU times required in the sample calculations described above are shown in Fig. 2 as functions of molecular length. These computations were performed with 200-decimal-digit accuracy on one R10000 64-bit processor of a Silicon Graphics Power Challenge computer. These execution times grow quadratically with molecular length N, as expected.

As noted above, lower precision calculations suffice for small molecules, while higher precision is required to treat larger ones. If the run time for a particular problem depended strongly on the required precision, an additional scaling with

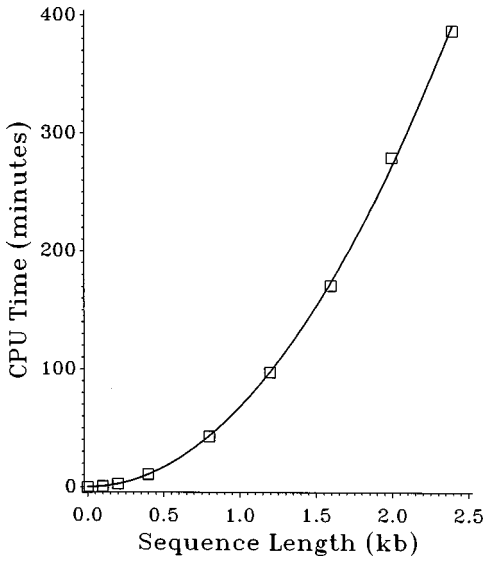

FIG. 2. The CPU time required for an exact calculation using Bailey's MPFUN high precision FORTRAN package [49] with 200 decimal digits of accuracy is plotted as a function of molecular sequence length (squares). These calculations were performed on a single R10000 64-bit RISC-based Silicon Graphics processor. The curve gives the best quadratic fit to the data points.

system size N could be introduced. However, the run time for calculations of up to 1000 decimal digits of accuracy in the multiprecision implementation of Bailey used here is dominated by the overhead from invoking the arbitrary precision subroutines, so that execution speed does not depend significantly on the precision specified [49].

Accelerated algorithm

Under typical physiological conditions the most frequently occupied states of superhelical DNA denaturation will have a relatively small number of open base pairs. We have taken advantage of this fact to develop a modified algorithm that retains a prescribed degree of accuracy at greatly reduced computational cost. In essence, this confines attention in a controlled way to the terms that dominate the partition function.

Combining Eq. (24) for the partition function Z with Eq. (25) for $Q(n)$, rearranging terms, and dropping the subscript λ referring to different treatments of the twist, we obtain

$$Z=\sum_{m=0}^{N} Z(m)=\sum_{m=0}^{N} Q(m)\mathcal{Z}(m), \tag{97}$$

where

$$\mathcal{Z}(m)=\sum_{S} \delta_{m,n}\exp\left\{-\beta\sum_{j=1}^{N}[(a+b_j)n_j-an_jn_{j+1}]\right\}. \tag{98}$$

The index m in Eqs. (97) and (98) refers to the number of open base pairs in the states being considered. For each m the set of states summed over in Eq. (98) are those for which the total number of open base pairs is $\Sigma_{j=1}^{N}n_j=m$. In what follows we denote by $Z^{(M)}$ the partial sum of the terms in the partition function up to and including $m=M$:

$$Z^{(M)}=\sum_{m=0}^{M} Z(m). \tag{99}$$

To find an upper bound M that suffices to guarantee a specified level of accuracy we proceed as follows. For each value of m, we find a lower bound $Z_L(m)$ and an upper bound $Z_U(m)$ for the term $Z(m)$:

$$0<Z_L(m)\leqslant Z(m)\leqslant Z_U(m), \quad 1\leqslant m\leqslant N. \tag{100}$$

Then

$$Z-Z^{(M)}=\sum_{m=M+1}^{N} Z(m)\leqslant \sum_{m=M+1}^{N} Z_U(m)=Z_U-Z_U^{(M)}, \tag{101}$$

where $Z_U^{(M)}=\Sigma_{m=0}^{M}Z_U(m)$ and $Z_U=\Sigma_{m=0}^{N}Z_U(m)$. Also,

$$Z_L=\sum_{m=0}^{N} Z_L(m)\leqslant Z. \tag{102}$$

Hence,

$$\frac{Z-Z^{(M)}}{Z}\leqslant\frac{Z_U-Z_U^{(M)}}{Z_L}. \tag{103}$$

The expression on the right hand side of this inequality is an upper bound on the fractional accuracy that is sacrificed when the true partition function Z is replaced by $Z^{(M)}$, which is the value obtained when states having more than M simultaneously open base pairs are ignored. Suppose that, for a specified ϵ, we can find an M such that

$$\frac{Z_U-Z_U^{(M)}}{Z_L}<\epsilon. \tag{104}$$

Then Eq. (103) shows that this ϵ also bounds the error arising when Z is approximated by $Z^{(M)}$:

$$\frac{Z-Z^{(M)}}{Z}<\epsilon. \tag{105}$$

We note that the same bound also applies to calculations of other quantities needed to evaluate ensemble averages. For example, consider the quantity $Z(n_l)$ that is defined in Eq. (44) and used to calculate the probability p_l that base pair l is separated. Denote by $\{Z(n_l)\}(m)$ the contribution to $Z(n_l)$ from all states with m open base pairs. Because $0\leqslant n_l\leqslant 1$, $\{Z(n_l)\}(m)\leqslant Z(m)$ for all m. Therefore

$$Z(n_l)-\sum_{m=1}^{M}\{Z(n_l)\}(m)=\sum_{m=M+1}^{N}\{Z(n_l)\}(m)$$
$$\leqslant\sum_{m=M+1}^{N} Z(m)=Z-Z^{(M)}<\epsilon Z. \tag{106}$$

So the approximation involved in ignoring states with more than M open base pairs will result in an error in the calcula-

tion of each p_l which is less than ϵ. Similar reasoning applies to $Z(n_l,n_k)$ and the other quantities used to calculate ensemble averages.

Upper and lower bounds may be calculated as follows. Let $b_{\min}$ and $b_{\max}$ denote the minimum and the maximum values of the separation energy parameter b_i over the sequence being considered, and assume $b_{\min},b_{\max}>0$. (In the copolymer analysis, $b_{\min}$ and $b_{\max}$ would be b_{AT} and b_{GC}, respectively.) One determines a lower bound on the energy of every state if one ascribes the value $b_{\min}$ to each open base pair. Because the energies enter the partition function with a negative exponent, this will provide an upper bound $Z_U(m) \geqslant Z(m)$ for every m. Similarly, ascribing the energy $b_{\max}$ to every open base pair yields a lower bound $Z_L(m) \leqslant Z(m)$. One can easily enumerate the contributions to the partition function from all states with m open base pairs under circumstances where all base pairs have the same transition energy b (i.e., the transition is homopolymeric). Specifically, one has $(m \neq 0)$

$$Z(m)=Q(m)e^{-\beta bm}\sum_{r=1}^{r_{\max}} \mathcal{N}(m,r)e^{-\beta ar}, \tag{107}$$

where the maximum number of open regions $r_{\max}=m$ if $m \leqslant N/2$, and $r_{\max}=N-m$ otherwise. (Recall that the molecule under consideration is circular.) Here $\mathcal{N}(m,r)$ is the number of states having m open base pairs in r runs, which is [5]

$$\mathcal{N}(m,r)=\sum_{r=1}^{r_{\max}} \frac{N}{r}\binom{m-1}{r-1}\binom{N-m-1}{r-1}. \tag{108}$$

Therefore, $Z_L(m)$ and $Z_U(m)$ both have the form

$$Z(m)=Q(m)e^{-\beta bm}\sum_{r=1}^{r_{\max}} \frac{N}{r}\binom{m-1}{r-1}\binom{N-m-1}{r-1}e^{-\beta ar}, \tag{109}$$

where $b=b_{\min}$ when calculating $Z_U(m)$, and $b=b_{\max}$ when calculating $Z_L(m)$.

One may then compute easily calculable bounds on Z_L and Z_U by considering bounds on the sum

$$\sum_{r=1}^{r_{\max}} \frac{e^{-\beta ar}}{r}\binom{m-1}{r-1}\binom{N-m-1}{r-1}=\sum_{r=1}^{r_{\max}} f_m(r). \tag{110}$$

The ratio of successive terms in this sum is

$$\frac{f_m(r+1)}{f_m(r)}=e^{-\beta a}\frac{(m-r)(N-m-r)}{r(r+1)}, \tag{111}$$

which is monotonically decreasing with r. For given m and N, the $r=1$ term $f_m(1)$ is the largest of the $f_m(r)$'s when

$$(m-1)(N-m-1)<2e^{\beta a}. \tag{112}$$

This will be true for all m when it holds for the m which makes the left hand side of this inequality largest, which is $m=N/2$. At $T=310\,^{\circ}\mathrm{K}$ with $a=10.84$ kcal/mol, one finds that the above inequality holds for all m whenever $N<18\,700$ base pairs. Because the ratio $f_m(r+1)/f_m(r)$ is monotonically decreasing, its largest value is $\rho_m=f_m(2)/f_m(1)$. This yields the following bounds:

$$\sum_{r=1}^{r_{\max}} f_m(r)<f_m(1)[1+\rho_m+\rho_m^2+\cdots]=\frac{f_m(1)}{1-\rho_m}. \tag{113}$$

Moreover, when $e^{-\beta a}(N-m-1)(m-1)<1$, one has $\rho_m<0.5$ for all m, so that the upper bound does not exceed $2f_m(1)$. Under the above posited conditions this occurs for all m whenever $N<13\,200$ base pairs. Also, $f_m(1)=e^{-\beta a}$, independent of m. Insertion of these bounds into Eqs. (101) and (102) gives the upper bound

$$Z-Z^{(M)}<2Ne^{-\beta a}\sum_{m=M+1}^{N} Q(m)e^{-\beta b_{\min}m}=B_U(M) \tag{114}$$

and the lower bound

$$Z>e^{-\beta K\alpha^2/2}+Ne^{-\beta a}\sum_{m=1}^{N} Q(m)e^{-\beta b_{\max}m}=B_L. \tag{115}$$

It requires $O(N)$ operations in total to calculate B_L and all the upper bounds $B_U(M)$ $(0 \leqslant M \leqslant N)$. If we can find an M for which $B_U(M)/B_L<\epsilon$, then Eqs. (103), (114), and (115) together show that ϵ also provides an upper bound on the error that arises when one disregards states with more than M open base pairs. This simplification reduces the operation count from $O(N^2)$ to $O(MN)$, which in practice can reduce the computational time by an order of magnitude or more. For example, consider the pBR322 DNA molecule containing 4363 base pairs, short enough for the above bounds to be valid under physiological conditions. This analysis guarantees that an accuracy of at least $\epsilon=10^{-9}$ will be achieved when $M=442$. In practice the actual accuracy may greatly exceed that suggested by the above estimate.

We also note that, in the accelerated algorithm, *both* the summation variables m and k in Eqs. (29)–(33) now take on only the values $[0,\ldots,M]$ rather than $[0,\ldots,N]$. Hence, for a given temperature and base pair sequence, each calculation at an additional linking difference requires only the smaller $O(M^2)$ incremental computational time.

Performance of the accelerated algorithm

Sample calculations were performed to evaluate the dependence on M of the execution time and accuracy achieved by this accelerated algorithm. The strand separation behavior was analyzed in pBR322 DNA ($N=4363$ base pairs) supercoiled to a linking difference of $\alpha=-27$ turns. This is the linking difference the molecule is found to have, on average, when it is extracted from bacteria. The energy parameters were assigned the values that apply under the conditions of the nuclease digestion procedure by which superhelical strand separation is experimentally detected [16,49].

The analysis was performed using the accelerated algorithm with various values for the upper bound M in the range $100 \leqslant M \leqslant 200$ base pairs. For comparison this transition was also analyzed using the complete exact algorithm (i.e., $M=N$). In both cases the probability p_l of strand separation

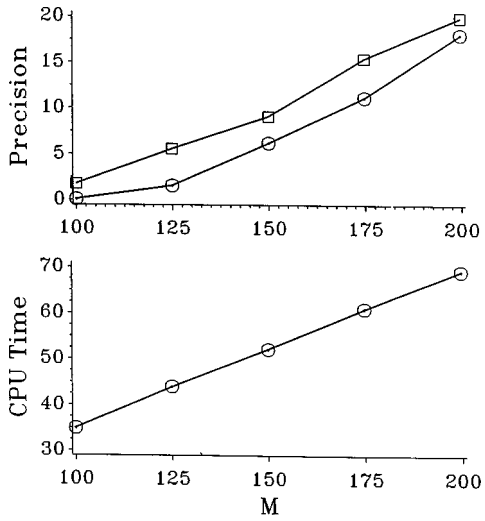

FIG. 3. The top graph plots the average accuracy (squares) and the minimum accuracy (circles) achieved by the accelerated algorithm as functions of the bound M imposed on the number of denatured base pairs in a state. The bottom graph plots the execution time of the accelerated algorithm as a function of M, which is seen to be linear in M to high accuracy. These calculations were performed on the pBR322 DNA sequence referred to in the text, and implemented on a dedicated HP 9000/735 RISC-based work station.

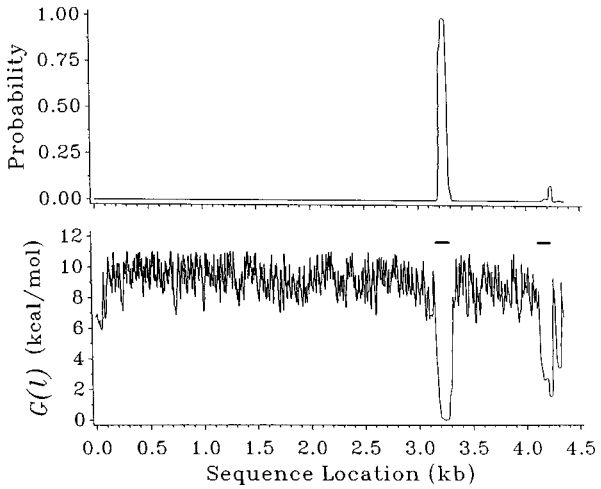

FIG. 4. The top graph shows the ensemble average probability of denaturation of each base pair in the pBR322 DNA molecule (4363 bp), on which a linking difference of $\alpha=-27$ turns is imposed. The exact algorithm was used in this calculation. The bottom graph plots the ''free energy'' of denaturation $\Delta G(l)$ vs position l, computed using Eq. (47). The sites where denaturation is experimentally observed to occur are denoted by bars. The predictions of these calculations are in precise quantitative accord with experimental observations, as described in the text.

was calculated for every base pair $1\leqslant l\leqslant N$. For each value of M the accuracy of each p_l was calculated as

$$A(l)=-\log_{10}|p_l-p_l(M)|, \tag{116}$$

the number of decimal digits of agreement with the exact value. [Here p_l and $p_l(M)$, respectively, denote the values calculated using the exact and accelerated algorithms, the latter with threshold M.] The accuracy of the accelerated algorithm was evaluated by finding both the minimum value of $A(l)$ over the entire sequence and its average value. Figure 3 plots these two values as functions of the bound M. These results show that the accelerated algorithm achieves very high accuracy at values of M considerably smaller than those estimated in the previous section. More than 18 decimal digits of accuracy are achieved when $M=200$, which is less than 5% of the number of base pairs in this molecule.

Figure 3 also displays the dependence of execution time on M when the calculations are performed on an HP 9000/735 computer, which has a RISC-based 32-bit processor. The CPU time required to perform these calculations is seen to increase linearly with M to high accuracy.

The above results show that under reasonable conditions one can retain a very high degree of accuracy with a moderate value of $M\approx 0.05N$, reducing the required computation time by an order of magnitude or more. The value of M can be selected using formulas giving rigorous bounds, as was done in the preceding section, or M can be estimated and the accuracy checked by comparing simulation results for different values of M. We find that simulations analyzing DNA molecules having lengths of biological interest (several kilobase pairs) typically require a few hours on a high-end workstation, which makes the accelerated algorithm a practical method for calculating superhelical strand separation behavior under a wide variety of circumstances.

COMPARISONS WITH EXPERIMENTAL RESULTS

In vitro experiments

The locations and extents of *in vitro* superhelical denaturation can be determined experimentally using the mung bean nuclease digestion procedure developed by Kowalski, Natale, and Eddy [48]. This enzyme cuts single strands of DNA but does not cut the duplex. Sites of cutting may then be located by sequencing, and the relative frequencies of cutting at different locations may be determined. The most detailed experimental analysis of superhelical denaturation applied these procedures to the pBR322 DNA molecule ($N=4363$ bp) [48].

Sample calculations have been performed on the pBR322 DNA molecule using the exact algorithm developed above. The linking difference was chosen to be $\alpha=-27$ turns, consistent with physiological values. Copolymeric transition energies were assumed, which ascribe one value b_{AT} to each AT base pair, and another value b_{GC} to each GC base pair. These and all other energy parameters were assigned values that were previously shown to be accurate under Kowalski and Eddy's experimental conditions [16]. Figure 4 shows the results of these calculations. The top portion of Fig. 4 gives the computed transition profile, the graph of the ensemble average probability p_l of denaturation versus position l. The bottom graph depicts the variation of the destabilization free energy $\Delta G(l)$ with position, as calculated from p_l using Eq. (47). The bars denote the locations where denaturation has been experimentally determined to occur [48].

Denaturation under these conditions was found experimentally to be confined to two locations. The primary location is between positions 3181 and 3300, coincident with the

terminator of the β-lactamase gene. The secondary location is between positions 4130 and 4250, at the promoter of the same gene. The amount of denaturation detected at the secondary site was 7% of that found at the primary site. The predictions of the exact method, like those of the previously developed approximate method, are in precise quantitative agreement with these experimental results. Transition is predicted to be confined to these two sites. Moreover, the areas under the transition probability curve in these regions, which give the expected number of denatured base pairs in each, agree with the relative amounts of denaturation experimentally observed there. This shows that the exact method developed here, applied to the model of Eq. (17) with no adjustable parameters, can provide quantitatively correct predictions of the denaturation behavior of DNA under the conditions of the nuclease digestion procedure by which it is experimentally detected [48]. It also suggests that the sites that are destabilized by superhelical stresses do not occur at random, but rather coincide with specific regulatory regions.

Comparison with *in vivo* results

Calculations performed using a previously developed, approximate, method have demonstrated close associations between stress-destabilized sites and several specific classes of regulatory regions [1,28]. Experiments have established that stress-induced denaturation does occur *in vivo* [19]. Here we assess the accuracy of the presently developed technique at predicting *in vivo* stress-induced denaturation by analyzing the 4-kb region containing the yeast FBP1 gene sequence used in these experiments. We note that this region is not circular *in vivo*. However, because the superhelical constraint of imposed linking difference is functionally identical in circular and in looped domains, one can treat regions that are not circular by a simple modification of the approach presented above. One simply conceptually closes the region into a circle by connecting its ends with a short run of *GC* base pairs, and then imposes a linking difference on the resulting domain. If the region were circularized by directly joining its ends instead of connecting them with an insert, one would run the risk of creating a spurious susceptible site in cases where the ends were reasonably $A+T$ rich. A $G+C$-rich insert is chosen because it has a high energy of denaturation, and hence will join the ends without either assisting their destabilization or itself constituting an introduced destabilized site. [Alternatively, one can join the two ends with a base pair which is constrained to remain closed (bonded).]

The results of this analysis, applied to the model of Eq. (17), are shown in Fig. 5. Here the linking difference selected is that which gives the level of torsional stress found in extracted plasmids. The bar denotes the region where denaturation was experimentally detected *in vivo*. One sees that this analysis provides a quantitatively precise depiction of the denaturation experienced in this region, even though the *in vivo* conditions are much more complex than those envisioned in the present model.

To determine what effect the alternative methods of treating the twists τ_j might have on the results, sample calculations using each assumption were performed on this sequence. In case (1) the τ_j's are assumed all to have the same value τ, which equilibrates with the residual superhelicity α_r. This was the assumption made in the previously developed, approximate method [5]. In case (2) the τ_j's can fluctuate independently. Figure 6 shows the profiles calculated in these two cases. The differences between the two profiles are quite slight, and are confined to the boundaries of the denaturing region.

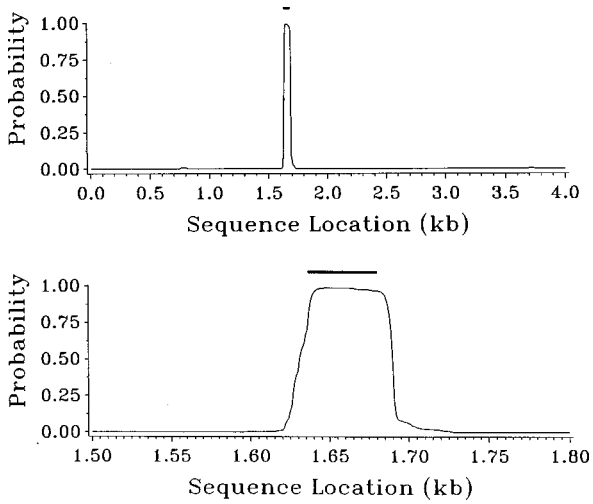

FIG. 5. The top graph shows the ensemble average probability of denaturation of each base pair in the 4 kb region of yeast DNA containing the FBP1 gene sequence. A linking difference of -22 turns was assumed, which gives the level of stress commonly found in DNA that has been extracted from living cells. The lower graph shows a detailed view of the region where denaturation is predicted to occur. In both graphs the bar indicates the region where denaturation is experimentally detected.

All the locations that denature in the two sequences analyzed in this section serve important regulatory functions. In the pBR322 plasmid denaturation is confined to two regions, the terminator and the promoter of the amp^r gene. In the yeast sequence denaturation occurs only at the terminal region of the FBP1 gene. These and many other results show that sites of stress-induced destabilization are closely associated with several types of regulatory regions [1,28,21]. This suggests that the interplay between base sequence and torsional stress provides a biologically important mechanism for regulatory activity.

DISCUSSION

This paper presents a method for calculating equilibrium local denaturation (strand separation) properties of superhelical DNA having kilobase lengths and specified sequences. The effective Hamiltonian includes the energies of denaturation of the *AT* and *GC* base pairs, the energies associated with the torsional deformations of the denatured regions, and interactions between denaturation and torsional deformations induced by the topological constraint of constant linking number L. The partition function and ensemble averages are calculated in a formally exact manner from the states of the system and their given energies. Through the introduction of auxiliary variables, the calculations can be implemented by an algorithm that is stable and has quadratic complexity with molecular length N. The computations can require relatively long execution times, however, since they must in general be implemented in arbitrary precision. An accelerated algorithm

was also developed which treats only states in which fewer than M base pairs are separated, and rigorous bounds on the resulting systematic error were established. This alternative method has $O(MN)$ complexity, although it still must be implemented in multiprecision arithmetic. It typically gives highly accurate results with modest choices of M ($M \approx 0.05N$), and can execute a realistic problem in less than one hour on a fast RISC-based work station. Calculations for additional linking numbers L can be obtained at a low incremental cost, $O(M^2)$ for each.

The methods developed here can be applied either to circular or, by a simple modification, to looped domains which are not topologically circular. Comparison with experiments indicates that this method, with no adjustable parameters, can provide quantitatively precise predictions for the locations and extents of superhelical DNA denaturation, both *in vitro* and *in vivo*.

Our approach has numerous advantages over the other techniques that have been previously used to treat superhelical denaturation. First, it correctly includes the topological constraint imposed by the domain structure, be it circular or looped, which is the fixing of the linking number L. Second, it permits the two inherently flexible single strands comprising a separated region to twist around each other. In this paper we present two levels of detail with which this twisting can be treated—either as a mechanical equilibrium torsional deformation or as a locally fluctuating quantity. Sample calculations show that these two alternatives give very similar results. Third, the approach correctly includes the exact contributions from all states, weighted according to their Boltzmann factors. As this is the only method able to evaluate the exact equilibrium distribution in polynomial time, it clearly improves on other techniques which either calculate an approximation of, or only sample from, that distribution.

This exact method can also treat many types of interesting cases that the other approximate methods, as presently formulated, cannot handle. These include near-neighbor base pair identity effects and alterations in the energies of base pair separation that result from such events as base methylation, protein or ligand binding, or the presence of pyrimidine dimers or abasic sites. Structural alterations that decrease the energies of denaturation of the base pairs involved have been predicted to significantly affect the transition behavior of stressed DNA [50]. The method can treat cases where one or more of the base pairs is externally constrained to remain open or closed, and is applicable to transitions at any temperature.

The approach presented in this paper can easily be extended to include the possibility of other competing transitions, such as cruciform extrusion at inverted repeat sequences, or *Z*-DNA or *H*-DNA formation at sites having the local base sequences required by these alternate conformations. An analysis in which c different conformations compete requires $c \times c$ transfer matrices. However, at present there is no compelling experimental evidence to suggest that transitions to any DNA conformations other than local denaturation serve biological functions; and, in any case, the energetics of these alternative transitions are not now known with sufficient precision to enable quantitatively accurate predictions of multistate competing transitions to be made.

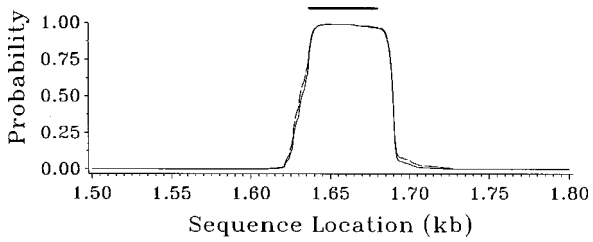

FIG. 6. Sample calculations were performed on the yeast FBP1 gene region to determine the effect of the different ways used to treat the twists τ_j in the denatured regions. In case (1), all τ_j's have the same value, the one which minimizes the Hamiltonian. In case (2), the τ_j's are allowed to fluctuate independently. The results from case (1) are plotted with a solid line, and those from case (2) with a dashed line. There are very slight differences between the two results, which are confined to the edges of the denatured region.

This approach also can be easily extended to include the possibility of a sequence-dependent nucleation energy a [32]. This could be important, for example, when analyzing the effects of the presence of abasic sites on transitions.

Future work will include the application of this method to a variety of DNA sequences for comparison with experimental data. The possibly important effects of various types and locations of defects, as mentioned above, will also be explored.

ACKNOWLEDGMENTS

The work of R.M.F. was supported by the U.S. DOE MICS Program under Contract No. DE-ACO4-94AL8500. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. DOE. The work of C.J.B. was supported by Grant No. BIR 93-10252 from the National Science Foundation and Grant No. RO1-GM47012 from the National Institutes of Health. We would like to acknowledge the hospitality of the Aspen Center for Physics, where part of this work was performed. R.M.F. would also like to acknowledge helpful conversations with R. Allen, D. Day, J. Delaurentis, and B. Hendrickson. We are especially grateful for the assistance of David Bailey in implementing an arbitrary precision version of the algorithm.

[1] C. J. Benham, J. Mol. Biol. **255**, 425 (1996).
[2] B. Alberts, D. Bray, J. Lewis, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell* (Garland, New York, 1994).
[3] M. Gellert, R. Menzel, K. Mizuuchi, M. O'Dea, and D. Friedman, Cold Spring Harbor Symp. Quant. Biol. **47**, 763 (1982).
[4] Z. Wang and P. Droege, J. Mol. Biol. **271**, 499 (1997).
[5] C. J. Benham, J. Chem. Phys. **92**, 6294 (1990).
[6] J. Marmur and P. Doty, J. Mol. Biol. **5**, 109 (1962).
[7] K. Breslauer, R. Frank, H. Bloecker, and L. Marky, Proc. Natl. Acad. Sci. USA **83**, 3746 (1986).

[8] S. G. Delacourt and R. D. Blake, J. Biol. Chem. **266**, 15 160 (1991).
[9] G. Steger, Nucleic Acids Res. **22**, 2760 (1994).
[10] D. Poland and H. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic, New York, 1970).
[11] R. M. Wartell and A. S. Benight, Phys. Rep. **126**, 67 (1985).
[12] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[13] T. Dauxois, M. Peyrard, and A. R. Bishop, Phys. Rev. B **47**, R44 (1993).
[14] D. Cule and T. Hwa, Phys. Rev. Lett. **79**, 2375 (1997).
[15] D. Kowalski and M. Eddy, EMBO J. **8**, 4335 (1989).
[16] C. J. Benham, J. Mol. Biol. **225**, 835 (1992).
[17] G. Michelotti, E. Michelotti, A. Pullner, R. Duncan, D. Eick, and D. Levens, Mol. Cell. Biol. **16**, 2656 (1996).
[18] S. D. Sheridan, C. J. Benham, and G. W. Hatfield, J. Biol. Chem. **273**, 21 298 (1998).
[19] A. Aranda, J. E. Perez-Ortin, C. J. Benham, and M. del Olmo, Yeast **13**, 313 (1997).
[20] J. Bode, T. Schlake, M. Rios-Ramirez, C. Mielke, M. Stengert, V. Kay, and D. Klehr-Wirth, Intl. Rev. Cytol. **162**, 389 (1995).
[21] C. J. Benham, T. Kohwi-Shigematsu, and J. Bode, J. Mol. Biol. **274**, 181 (1997).
[22] M. Tal, F. Shimron, and G. Yagil, J. Mol. Biol. **243**, 179 (1994).
[23] C. J. Benham, Proc. Natl. Acad. Sci. USA **76**, 3870 (1979).
[24] A. Anshelevich, A. Vologodskii, and M. Frank-Kamenetskii, J. Biomol. Struct. Dyn. **6**, 247 (1988).
[25] V. Anshelevich, A. Vologodskii, A. Lukashin, and M. Frank-Kamenetskii, Biopolymers **18**, 2733 (1979).
[26] V. Bloomfield, D. Crothers, and I. Tinoco, *Physical Chemistry of Nucleic Acids* (Harper and Row, New York, 1974).
[27] S. Katsura, F. Makishima, and H. Nishimura, J. Biomol. Struct. Dyn. **10**, 639 (1993).
[28] C. J. Benham, Proc. Natl. Acad. Sci. USA **90**, 2999 (1993).
[29] J. D. Engel and P. H. von Hippel, J. Biol. Chem. **253**, 927 (1978).
[30] M. Clooins and R. M. Myers, J. Mol. Biol. **198**, 737 (1987).
[31] A. I. H. Murchie and D. M. J. Lilley, J. Mol. Biol. **205**, 593 (1989).
[32] G. Vesnaver, C. N. Chang, M. Eisenberg, A. P. Grollman, and K. J. Breslauer, Proc. Natl. Acad. Sci. USA **86**, 3614 (1989).
[33] C. A. Gelfand, G. E. Plum, A. P. Grollman, F. Johnson, and K. J. Breslauer, Biochemistry **37**, 12 507 (1998).
[34] H. Sun, M. Mezei, R. Fye, and C. J. Benham, J. Chem. Phys. **103**, 8653 (1995).
[35] Note that we implicitly include the term ''pV'' in H_0, and V is assumed to be integrated over, in order to obtain the correct partition function for the experimentally relevant (constant temperature and pressure) Gibbs free energy.
[36] J. C. Wang, Proc. Natl. Acad. Sci. USA **76**, 200 (1979).
[37] C. Schildkraut and S. Lifson, Biopolymers **3**, 195 (1968).
[38] H. Yamaki, E. Ohtsubo, K. Nagai, and Y. Maeda, Nucleic Acids Res. **16**, 5067 (1988).
[39] A. L. Oliver, R. M. Wartell, and R. L. Ratliff, Biopolymers **16**, 1115 (1977).
[40] B. Amirikyan, A. Vologodskii, and Y. Lyubchenko, Nucleic Acids Res. **9**, 5469 (1981).
[41] A. I. H. Murchie, R. Bowater, F. Aboul-ela, and D. M. J. Lilley, Biochim. Biophys. Acta **1131**, 1 (1992).
[42] W. R. Bauer and C. J. Benham, J. Mol. Biol. **234**, 1184 (1993).
[43] W. R. Bauer and J. Vinograd, J. Mol. Biol. **47**, 419 (1970).
[44] R. Depew and J. C. Wang, Proc. Natl. Acad. Sci. USA **72**, 4275 (1975).
[45] D. Pulleyblank, M. Shure, D. Tang, J. Vinograd, and H.-P. Vosberg, Proc. Natl. Acad. Sci. USA **72**, 4280 (1975).
[46] H. A. Kramers and G. H. Wannier, Phys. Rev. **60**, 252 (1941).
[47] F. B. Fuller, Proc. Natl. Acad. Sci. USA **75**, 3557 (1978).
[48] D. Kowalski, D. Natale, and M. Eddy, Proc. Natl. Acad. Sci. USA **85**, 9464 (1988).
[49] D. H. Bailey, ACM Trans. Math. Softw. **19**, 288 (1993).
[50] C. J. Benham, J. Mol. Biol. **150**, 43 (1981).