JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 4, 2004

Pp. 519–543

The Analysis of Stress-Induced Duplex Destabilization in Long Genomic DNA Sequences

CRAIG J. BENHAM and CHENGPENG BI

ABSTRACT

We present a method for calculating predicted locations and extents of stress-induced DNA duplex destabilization (SIDD) as functions of base sequence and stress level in long DNA molecules. The base pair denaturation energies are assigned individually, so the influences of near neighbors, methylated bases, adducts, or lesions can be included. Sample calculations indicate that copolymeric energetics give results that are close to those derived when full near-neighbor energetics are used; small but potentially informative differences occur only in the calculated SIDD properties of moderately destabilized regions. The method presented here for analyzing long sequences calculates the destabilization properties within windows of fixed length N, with successive windows displaced by an offset distance d_o. The final values of the relevant destabilization parameters for each base pair are calculated as weighted averages of the values computed for each window in which that base pair appears. This approach implicitly assumes that the strength of the direct coupling between remote base pairs that is induced by the imposed stress attenuates with their separation distance. This strategy enables calculations of the destabilization properties of DNA sequences of any length, up to and including complete chromosomes. We illustrate its utility by calculating the destabilization properties of the entire *E. coli* genomic DNA sequence. A preliminary analysis of the results shows that promoters are associated with SIDD regions in a highly statistically significant manner, suggesting that SIDD attributes may prove useful in the computational prediction of promoter locations in prokaryotes.

Key words: DNA structural transitions, DNA supercoiling, genome annotation.

I. INTRODUCTION

THE DNA DOUBLE HELIX IS NOT A PERMANENT STRUCTURE. During its physiological activities, circumstances arise that require the two strands of the DNA duplex to become transiently separated. Local strand separation within promoters is an obligatory step in the initiation of transcription, enabling the RNA polymerase to gain access to the bases in the template strand of the encoding DNA. The initiation of DNA replication also requires strand separation so the bases within each strand may serve as templates for the synthesis of their respective complements. Because initiation of the two central activities of DNA both require local separation of the DNA duplex, the locations and occasions where strand separations occur must be stringently regulated *in vivo*.

UC Davis Genome Center, University of California, One Shields Avenue, Davis, CA 95616.

Regulatory DNA strand separation commonly involves interactions with other molecules such as DNA binding proteins, polymerase complexes and σ-factors, topoisomerase or helicase enzymes, transcription factors, and other activators and/or inhibitors. Numerous examples are known of regulatory proteins that bind DNA in a single strand-specific manner (Rothman-Denes *et al.*, 1998). In some cases, these molecules do not participate in separating the DNA duplex to create the single strands to which they bind, but instead require a preexisting separated region. The FBP protein, whose binding regulates *c-myc* oncogene transcription, is an example of this class (He *et al.*, 2000). Although other single strand-specific DNA binding proteins may actively contribute to opening the double helix, the DNA duplex still may need to be partially or entirely destabilized at the separation site for this opening to occur efficiently.

One biologically important way in which the local stability of the DNA double helix is regulated is through superhelical stresses imposed on the DNA duplex (Benham, 1979). These stresses are modulated *in vivo* by a variety of processes, including gyrase and topoisomerase activities, translocation of RNA polymerase during transcription, changes of nucleosome binding patterns, activities of DNA helicases, or constraints imposed by other DNA binding events. The transition from stationary phase to growth phase in *E. coli* induces a rapid increase in the level of negative superhelicity of the entire chromosome. In response, the global pattern of gene expression changes, allowing the organism to rapidly adjust to the higher metabolic demands of growth (Hatfield and Benham, 2002). Environmental changes, such as the transition to anaerobicity or osmotic shock, also induce changes of chromosomal superhelicity in this organism. These attune its global expression patterns to its altered environment (Cheung *et al.*, 2003). Translocation of the RNA polymerase complex during transcription produces a bow wave of overtwist and leaves a wake of undertwist (Liu and Wang, 1987), which modulate the upstream and downstream levels of torsional stress accordingly.

Negative DNA superhelicity is the standard experimental way to impose stresses on DNA. Sufficient negative superhelicity can destabilize the DNA duplex, driving local denaturation as well as transitions to other alternative structures that are less twisted in the right-handed sense than is the B-form helix (Benham, 1981). Because strand separation transforms B-form DNA into an untwisted conformation, it concentrates some of the imposed negative superhelicity as a local decrease of twist. This relaxes by a corresponding amount the effective level of negative superhelicity experienced by the other base pairs in the domain. So the free energy cost of performing such a conformational transition in a superhelically stressed domain is partially or fully offset by the free energy returned from the fractional relaxation that results. Local sites of strand separation, the most extreme form of duplex destabilization, can be induced by moderate levels of negative superhelicity (Kowalski *et al.*, 1988). Partial destabilizations also can occur, in which the imposed superhelical stresses fractionally decrease the free energy needed to separate the duplex. Partial destabilization can be biologically important, as it decreases the amount of free energy other molecules must provide to drive separation at the site involved. This can greatly affect the occurrence, or the rate, of any reaction involving a single stranded DNA substrate.

The destabilization experienced by a negatively superhelical DNA sequence is not uniform, but instead is commonly highly concentrated at a relatively small number of positions where the intrinsic thermodynamic stability is small over a sufficiently long interval (Kowalski *et al.*, 1988). Thermodynamic stability is a strictly local attribute of the base sequence of DNA, depending in complex but experimentally well characterized ways on base pair identity, nearest neighbors, temperature, and ionic conditions (Steger, 1994; Breslauer *et al.*, 1986; Delacourt and Blake, 1991). A simple rule of thumb that is reasonably accurate under physiological conditions is that A+T-rich regions are, on average, less stable (i.e., have smaller free energies of denaturation) than other sites.

However, transitions driven by superhelical stresses within topologically constrained DNA domains behave very differently than do transitions driven by temperature in unconstrained molecules. The intrinsic propensity of a specific site to undergo a stress-induced transition depends on the difference between the energy cost of the transition and the energy benefit of the relaxation it affords. This relaxation alters the level of stress, and hence the probability of transition, in the rest of the domain. So transitions that involve only near-neighbor interactions when they occur in linear or nicked DNA, will in stressed DNA be coupled to the conformational states of all other base pairs experiencing the stress. Whether transition occurs at a given site depends not just on its local sequence properties (such as thermodynamic stability), but also on how that site competes with all others in the domain. In this way, superhelical stresses globally couple together the transition behaviors of all base pairs experiencing them. In consequence of this coupling,

stress-induced transitions have a large repertoire of highly intricate, nonlinear and interactive possible behaviors that far transcend what can occur in thermally driven transitions in unconstrained molecules (Benham, 1980, 1996a). Small changes in base sequence at one position can affect the probabilities of transition of other sites, potentially even many hundreds or even thousands of base pairs away. This is illustrated in Fig. 1, which shows that deletion of a short region can drastically alter the stress-induced duplex destabilization (SIDD) properties of an extensive region while changing the thermodynamic stability only at the deletion site. Further, as shown in Fig. 2, denaturation at one location can be coupled to the reversion back to B-form of remote denatured sites. In this example, the thermodynamic stabilities of the two sites involved are precisely equal. This complex coupled behavior contrasts with thermal denaturation, where the probability of transition of each base pair increases monotonically with temperature.

Because genomic DNA is topologically constrained *in vivo* and SIDD properties are determined by interactions among the base pairs experiencing a superhelical stress, no string-based surrogate attribute can accurately depict the propensity to denature of chromosomal DNA. In particular, scans of the thermodynamic stability of base sequences do not adequately illuminate this behavior (Benham, 1996a). The two sites which compete for destabilization in Fig. 2, for example, have the same level of thermodynamic stability.

DNA duplex destabilization in regulation

Stress-induced DNA duplex destabilization (SIDD) has been implicated in the mechanisms of activity of numerous biological processes. Activation of the *ilv*PG promoter of *E. coli* by integration host factor (IHF) involves a binding-induced transmission of superhelical destabilization from the upstream regulatory region to the −10 region (Sheridan *et al.*, 1998, 1999). In humans, initiation of transcription from the *c-myc* oncogene is regulated in part by binding of the FBP protein to the single stranded FUSE element (He *et al.*, 2000).

The initiation of eukaryotic transcription from relaxed templates requires transcription factors that are not needed when the substrate DNA is negatively supercoiled (Parvin *et al.*, 1994). Indeed, the minimal conditions for *in vitro* transcriptional initiation from the yeast *CUP1* promoter involve no other regulatory molecules at all; a negatively superhelical DNA substrate plus RNA polymerase (RNAP) was sufficient to induce transcription from this promoter (Leblanc *et al.*, 2000). Negative superhelicity destabilizes the DNA duplex in the region around the promoter, which enables RNAP to bind to the template strand. Once bound, RNAP found the transcription start site by a random walk-type search and initiated transcription there. This process did not require TATA box-binding protein, or any other factors.

Stress-induced DNA duplex destabilization (SIDD) also has been implicated in other transcriptional regulatory processes. Polyadenylation and transcription termination are coupled events in yeast, but not in higher eukaryotes. However, there is no clear consensus sequence associated with the sites within the 3' flanks of genes where these events occur. Analysis of the SIDD properties of 26 yeast genes showed that in every case the 3' flank was predicted to be strongly destabilized by stresses (Benham, 1996a). Subsequently, the FBP1 gene was placed in a plasmid that was inserted into yeast, and strand opening was experimentally assessed *in vivo* by S1 endonuclease digestion (Aranda *et al.*, 1997). The results showed that the site where destabilization was predicted to occur in fact was strand separated *in vivo* when this gene was active. *In vitro* deletion experiments showed that the site experiencing strand separation was the only region of the 3' flank whose presence was required for correct transcript termination.

The initiation of replication in both prokaryotes and yeast has been shown to require the presence at a precise position of a site which is susceptible to superhelical strand separation (Kowalski and Eddy, 1989; Huang and Kowalski, 1993). The base sequences of these stress-destabilized regions within *oriC* could be modified without affecting replication, provided they retained their susceptibility to stress-induced denaturation. However, changes that either decreased that susceptibility or moved the SIDD site by as little as 50 bp extinguished *in vivo* ORI function.

Recent work has shown that SIDD sites created by mutations can act as unintended replication origins. Specifically, spinocerebellar ataxia type 10 involves expansion of a pentameric DNA repeat. When *n*-mers of this repeat were placed in plasmids that did not contain replication origins, those that were sufficiently long to be associated with expansion *in vivo* both experienced stress-induced denaturation and activated plasmid replication. In contrast, *n*-mers that were too short for expansion did not undergo this transition

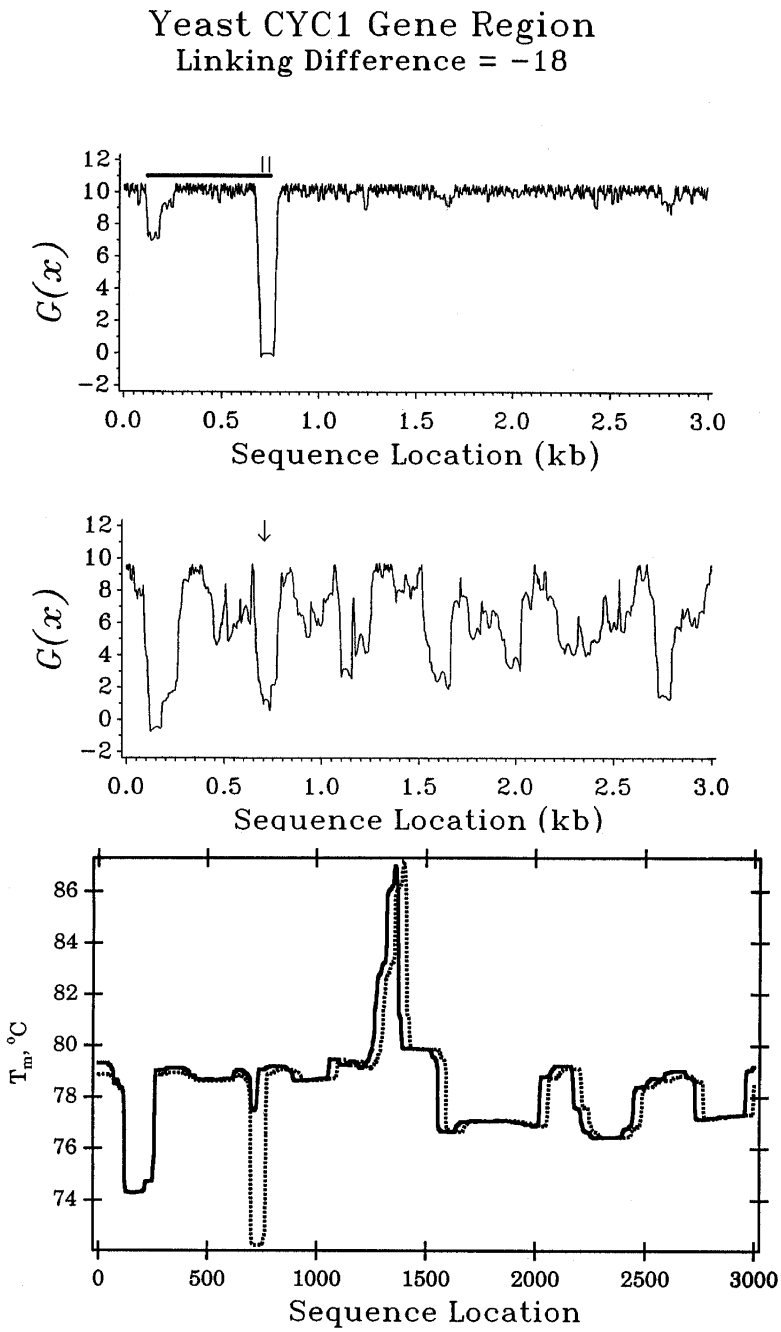

FIG. 1. The top graph plots the SIDD profile of a 3 kb genomic sequence from yeast that contains the CYC1 gene, whose transcribed region is denoted by the bar in the **upper** graph. The quantity plotted is $G(x)$, the incremental free energy needed to guarantee opening of the base pair at position x. High values of $G(x)$ correspond to stable positions, low values to unstable positions. (The method used to calculate this profile is described in Section II.) The wild-type sequence is largely stable, with destabilization confined to regulatory sites—the promoter and terminator of the gene. When the 38 bps located between the two vertical bars are deleted, the SIDD profile changes to that shown in the **middle** graph. Now the entire sequence is strongly and chaotically destabilized, with major changes extending even 2 kb away from the deletion site. The thermodynamic stability profiles of the wild-type (dotted line) and the mutated (solid line) sequences are presented in the **bottom** graph, which plots the transition temperature T_m at each position evaluated using the MELTMAP algorithm (Lerman and Silverstein, 1987). The deletion causes major changes of SIDD properties throughout the region, while altering the thermodynamic stability only at its site.

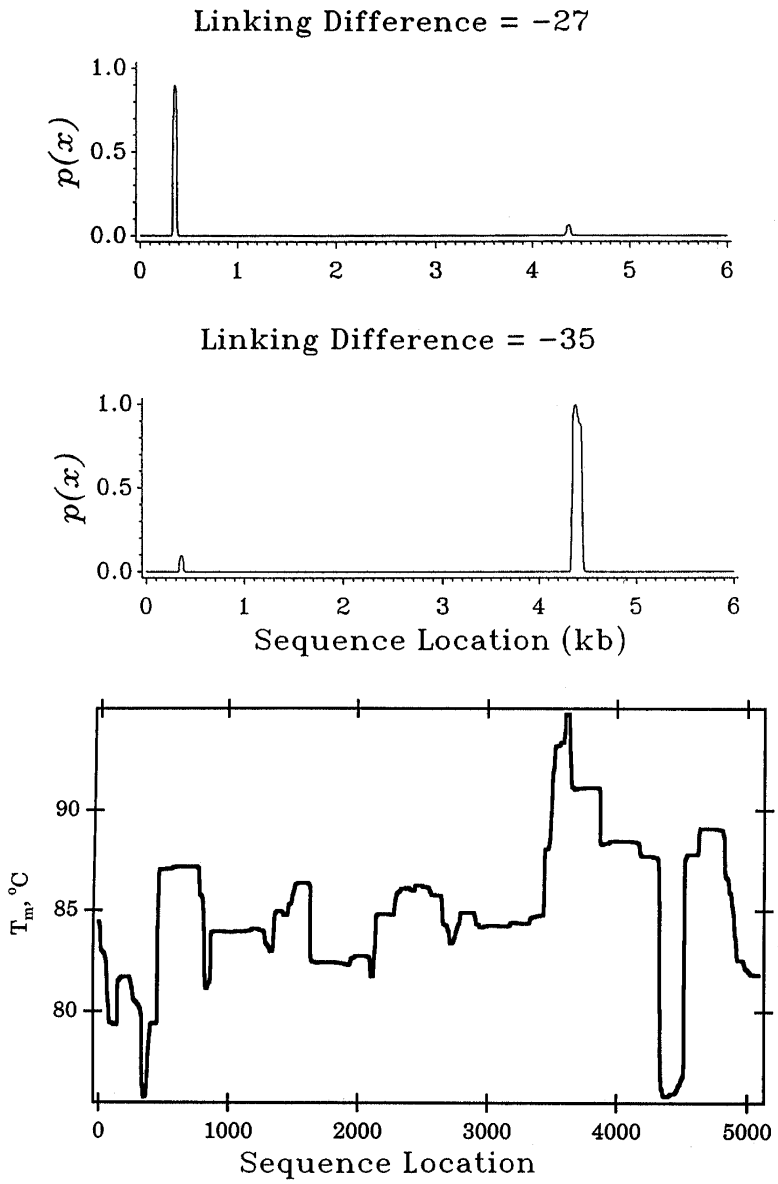

FIG. 2. The **top** two graphs show the denaturation probability profiles of a 6 kb region containing the hen histone H5 gene at two different levels of stress. The upper graph uses linking difference $\alpha = -27$ turns and the lower graph uses $\alpha = -35$ turns, which correspond to superhelix densities $\sigma = -0.046$ and $\sigma = -0.061$, respectively. Significant opening is confined to two sites, at positions 400 and 4,400. At the lower stress level the opening at position 400 dominates. However, at the higher stress level this site reverts back to B-form, coupled to the opening of the site at position 4,400. The **bottom** graph is the MELTMAP plot of the transition temperature profile of this region. One sees that the thermodynamic stabilities of these two competing sites are precisely equal, yet their stress-driven behaviors are complex, interactive, and nonlinear. The probability of transition at some positions does not increase monotonically with the denaturing constraint, as it does in thermal transitions in unconstrained molecules, but instead is coupled to the behaviors of other sites.

and did not induce replication (Potaman *et al.*, 2003). This suggests a SIDD-based mechanism by which the expansion characterizing this disease could occur.

Sites susceptible to stress-induced duplex destabilization also characterize a variety of chromosomal attachment regions. Examples include yeast centromeres (Tal *et al.*, 1994) and matrix attachment regions (MARs) which are positions where the eukaryotic chromosome putatively attaches to the interphase

nuclear matrix (Benham *et al.*, 1997). Although MARs occur at defined loci within the chromosomal DNA sequence, they do not display any well-characterized consensus sequence, so the basis for their positional specificity has not been clear. However, both experimental and theoretical evidence shows that they have characteristically strong propensities to denature under torsional stress (Bode *et al.*, 1992; Kohwi-Shigematsu and Kohwi, 1993; Benham *et al.*, 1997). This attribute may account for many of the regulatory effects MARs are known to exert on both transcription and replication.

Bizelesin is one of the most potent antitumor drugs known, lethal at dosages as small as ten molecules per cell (Woynarowski *et al.*, 2001). Bizelesin crosslinks the two strands of the duplex at A-A/T-A/T-A/T-A/T-A sites. Although the precise mechanism of bizelesin cytotoxicity has not been determined, it is possible that crosslinking keeps in duplex form a region that must locally separate for some essential biological process to occur. This may involve either MAR binding or the initiation of replication at ORIs.

These investigations have implicated stress-induced duplex destabilization (SIDD) in the regulation of a wide variety of biological processes. Most of these studies involved collaborations between experimental groups working on model systems, and this theoretical group, which performed computational analyses of the SIDD properties of the regulatory regions under investigation. Cumulatively, their results show that stress-induced destabilization is an essential component of many regulatory mechanisms governing a wide range of normal and pathological events. This underscores the importance of developing a computational method to assess the SIDD properties of long genomic sequences, including complete chromosomes. The information gleaned from such analyses would enable the investigation of associations between SIDD properties and every known type of regulatory region. In this paper, we present a method to perform these calculations.

The computational analysis of stress-induced DNA duplex destabilization

This research group has developed three methods to calculate the pattern of duplex destabilization experienced by a kilobase-long, circular DNA molecule in response to negative superhelicity (Benham, 1990, 1992; Sun *et al.*, 1995; Fye and Benham, 1999). These methods, which are described in Section II below, all evaluate the statistical mechanical equilibrium distribution among states of denaturation of a circular DNA molecule of specified base sequence, on which a defined level of superhelicity has been imposed. The conformational and free energy parameters that occur in these analyses are given their experimentally measured values, so there are no free parameters to be fit. Yet, in all cases where experiments have been performed, the computations were found to make quantitatively accurate predictions of the locations and extents of strand separation as functions of base sequence and imposed superhelicity (Benham, 1992). Moreover, many of the sites that were predicted to separate under stress have subsequently been experimentally shown to be destabilized, both *in vitro* and *in vivo*. This investigator's computational analyses of the SIDD properties of specific DNA sequences have played central roles in the collaborations which have elucidated many of the biological roles of stress-induced duplex destabilization.

However, as presently implemented these computational methods treat only relatively short DNA sequences, commonly less than 10 kb. It is clearly important to develop techniques to analyze the SIDD properties of much longer sequences. In this paper, we develop a technique for computing SIDD properties of complete chromosomal DNA sequences. We first present the most general and efficient implementation of the approximate method. This allows the user to assign denaturation energies individually to each base pair, which enables inclusion of near-neighbor effects, methylation, ligand binding, lesions, and other alterations that affect the local transition energy. Then we describe a windowing procedure for analyzing long sequences. Here, the SIDD profile of the sequence within an N-base-pair window is analyzed using the approximate method. Then the window is moved an offset distance d_o, and the sequence within this new window is analyzed. This procedure is repeated down the length of the molecule, so each base pair occurs in N/d_o windows. The final values of its SIDD properties are calculated as weighted averages of the properties computed for it in each window where it appears.

In this arrangement, the imposed stress tightly couples together the SIDD behaviors of closely spaced sites. But this direct coupling attenuates as the separation distance between sites increases, until sites separated by more than N base pairs are no longer *directly* coupled. However, each site is strongly coupled to its close neighbors, which in turn are coupled to their close neighbors, and so on in a chain of direct couplings that has the effect of *indirectly* coupling together more remote regions. This arrangement accords with biological expectations in that stresses within chromosomal DNA are unlikely to propagate globally, but instead to attenuate with distance.

II. THE METHOD OF SIDD ANALYSIS

Statistical mechanics of superhelical denaturation

DNA within a topological domain is constrained so that its linking number Lk is fixed, where Lk is the total number of helical turns within the domain when its central axis is held planar. The linking number of a relaxed domain is Lk_0. But DNA *in vivo* is commonly constrained in a negatively superhelical state, in which $Lk < Lk_0$. The resulting (negative) linking difference $\alpha = Lk - Lk_0 < 0$ exerts untwisting torsional stresses on the domain. If these stresses are sufficiently large, they can drive local structural transitions to conformations whose right-handed helicities are less than that of the B-form (Benham, 1981). Such transitions package some of the linking difference as the local change of twist they produce, which diminishes by a corresponding amount the negative torsional stress experienced by the balance of the domain. Although such a transition requires free energy, the fractional relaxation it produces in a topologically constrained domain also yields a free energy return. When that return exceeds its cost, the transition will be favored at equilibrium.

Local transitions that can be driven by negative DNA superhelicity include strand separation (also called local denaturation) (Vinograd *et al.*, 1968; Dean and Lebowitz, 1971; Kowalski *et al.*, 1988), cruciform extrusion at inverted repeat sequences (Woodworth-Gutai and Lebowitz, 1976; Lilley, 1980; Benham *et al.*, 2002), B-Z transitions (Peck *et al.*, 1982; Singleton *et al.*, 1982; Rahmouni and Wells, 1989), and B-H transitions (Kohwi-Shigematsu and Kohwi, 1991). Each of these superhelical transitions can be analyzed using the methods presented here. One need only change the geometric parameters describing the alternate state and the free energy parameters governing the transition, and confine the possibility of transition to those regions whose sequences permit it. We explicitly develop our method for strand separation because it is the only DNA structural transition that is presently known to be essential *in vivo*.

A given linking difference imposed on a DNA molecule can be accommodated by many combinations of superhelical deformations and conformational transitions. In particular, because every base pair can denature, there are 2^N states of strand separation in a domain containing N base pairs. At thermodynamic equilibrium, a population of identical superhelical DNA molecules will be distributed among these states, with high energy states exponentially less occupied than low energy states. If the states are indexed by i, and the free energy of state i is G_i, then the equilibrium probability p_i of state i, which is its fractional occupancy in a population at equilibrium, is

$$p_i = \frac{e^{(-G_i/RT)}}{\sum_i \exp^{(-G_i/RT)}},$$

where R is the gas constant, T is the absolute temperature, and the denominator is the partition function Z. (For simplicity, all states are denoted here as though they are discrete in character. For parameters that vary continuously, it is understood that relevant summations actually involve integrals.) If a parameter ζ has value ζ_i in state i, then its population average (i.e., expected) value $\bar{\zeta}$ at equilibrium is

$$\bar{\zeta} = \sum_i \zeta_i p_i.$$

This expression may be used to evaluate any property of interest, once the equilibrium distribution is known.

In our analysis, we calculate the ensemble average probability $p(x)$ of denaturation of the base pair at each position x along the DNA sequence. The graph of $p(x)$ versus x, called the transition profile, displays the regions of the sequence that have significant probabilities of denaturation. Figure 2 displays the calculated transition profiles of the hen H5 histone gene under two different levels of superhelicity.

A more sensitive measure of destabilization is given by the incremental free energy $G(x)$ needed to guarantee separation of the base pair at position x (Benham, 1993, 1996a). This $G(x)$ is the difference between the average free energy $\bar{G}(x)$ of all states in which base pair x is open and the average free energy $\bar{G}$ of the system. $G(x)$ is near zero or negative for base pairs that are favored to separate with high probability and positive for base pairs where incremental free energy is needed to assure separation.

Stress-induced duplex destabilization (SIDD) profiles are plots of $G(x)$ versus x. (Figure 1 shows SIDD profiles of a region and of a deletion mutant thereof.) SIDD profiles are more informative than transition profiles because they also depict sites where the amount of free energy needed to induce denaturation is decreased relative to background, but not to levels where opening becomes favored. This phenomenon will be important when duplex opening occurs by processes that can provide sufficient free energy to cause local denaturation only if the DNA site involved already is marginally destabilized by stresses. For example, decreasing the stability of a region by 2.4 kcal/mol will favor the open state at equilibrium by a factor of 100, even in a process that is modulated by protein interactions. Such regions could be biologically important as sites which stresses render vulnerable to opening by enzymatic or other processes.

The model of superhelical DNA denaturation

Here the imposed linking difference α is partitioned among three factors. First, denaturation of n base pairs in a domain decreases its total twist by n/A turns, where $A = 10.5$ base pairs per turn is the sequence-averaged helicity (twist per length) of unstressed B-form DNA (Wang, 1979). Torsional stresses will remain unless $n = -\alpha A$. Because single strands of DNA are much more flexible than is the B-form duplex (Bloomfield *et al.*, 1974), the separated strands in a denatured region will tend to twist around each other in response to residual stresses. We denote the total twist of the denatured regions by $\mathcal{T}$. Finally, the residual linking difference α_r is the component of α that is not accommodated by either of these two deformations. We need not decompose α_r further since, as discussed below, the free energy associated to it has been determined from experiments. The superhelical constraint coupling together these modes of accommodating a constant α is expressed in the conservation equation

$$\alpha = -\frac{n}{A} + \mathcal{T} + \alpha_r = \text{constant}. \tag{1}$$

Each of these DNA deformations has an associated free energy, which together determine the free energy associated to a state. First, the free energy of denaturation is

$$G_d = a\sum_{j=1}^{N} n_j(1-n_{j+1}) + \sum_{j=1}^{N} b_j n_j = \sum_{j=1}^{N}\Big\{(a+b_j)n_j - a n_j n_{j+1}\Big\}. \tag{2}$$

Here, the binary variable n_j assumes the value 0 if the base pair at position j is in the B-form and 1 if it is denatured. The first summation in Equation (2) counts the number of denatured regions in a circular domain, and the second refers to individual separated base pairs. Here, a is the nucleation free energy required to initiate a new denatured region. Its value has been measured experimentally to lie between 10 and 12 kcal/mol, depending on environmental conditions (Amirikyan *et al.*, 1981; Benham, 1992; Bauer and Benham, 1993; Bauer *et al.*, 1995). The term b_j is the free energy needed to denature the base pair at position j. Its value depends on the identity of this base pair, the identities of its neighbors, and environmental conditions. Both the entropy and enthalpy of denaturation have been measured to high accuracy for all ten types of neighbor base pairs at several ionic strengths (Steger, 1994). The free energy b_j also varies significantly for mispaired or unpaired bases (Benham *et al.*, 2002), for abasic sites (Vesnaver *et al.*, 1989), and for bases that have been chemically modified by methylation (Engels and von Hippel, 1974; Yamaki *et al.*, 1988; Murchie and Lilley, 1989), the formation of adducts (Gelfand *et al.*, 1998), or other types of molecular lesions.

The twist $\mathcal{T}$ associated with the denatured base pairs is

$$\mathcal{T} = \sum_{j=1}^{N} \frac{n_j \tau_j}{2\pi},$$

where τ_j (radians per base pair length) is the local helicity at each position j within the denatured regions. We associate a Hooke's Law torsional energy to this deformation

$$G_{\mathcal{T}} = \frac{C}{2}\sum_{j=1}^{N} n_j \tau_j^2,$$

which may be regarded as the lowest order term in a possibly more complex energy law. The effective torsional stiffness C associated to this interstrand twisting of the single strands in denatured DNA has been measured experimentally to be approximately 1/80 that of B-form DNA (Crothers and Spatz, 1971; Benham, 1992; Bauer and Benham, 1993). We do not explicitly consider twisting of the much stiffer B-form regions in response to stress, instead incorporating this deformation into the residual linking difference α_r.

Lastly, there is a free energy G_r associated with residual linking α_r. This has been determined from experiments to be quadratic in α_r to high accuracy,

$$G_r = \frac{K\alpha_r^2}{2} = \frac{K}{2}\left(\alpha + \frac{n}{A} - \mathcal{T}\right)^2,$$

and the coefficient K has been measured for numerous molecules at various temperatures and ionic strengths (Bauer and Vinograd, 1970; Depew and Wang, 1975; Pulleyblank *et al.*, 1975; Benham, 1992; Bauer and Benham, 1993).

To specify a state of this system, one must first specify the open base pairs, i.e., the values of the N binary variables n_j and the torsional deformation τ_j of each. This will determine the residual superhelicity through the conservation equation, Equation (1). The free energy associated to a state then is

$$G = G_{\mathcal{T}} + G_r + G_d = \frac{C}{2}\sum_{j=1}^{N} n_j\tau_j^2 + \frac{K}{2}\left(\alpha + \frac{n}{A} - \sum_{j=1}^{N}\frac{n_j\tau_j}{2\pi}\right)^2 + \sum_{j=1}^{N}\left\{(a+b_j)n_j - an_jn_{j+1}\right\}. \quad (3)$$

Because the values of all energy parameters have been measured experimentally, there are no free parameters in this analysis.

Statistical mechanical calculations

The partition function Z governing this system is

$$Z = \sum_{\mathcal{S}}\left\{Q(n)\exp\left(-\beta\sum_{j=1}^{N}\left\{(a+b_j)n_j - an_jn_{j+1}\right\}\right)\right\}, \quad (4)$$

where $\beta = 1/RT$,

$$Q(n) = \prod_{i=1}^{n}\int_{-\infty}^{\infty} d\tau_i \ \exp\left[-\beta\left\{\sum_{i=1}^{n}\frac{C\tau_i^2}{2} + \frac{K}{2}\left(\alpha + \frac{n}{A} - \sum_{i=1}^{n}\frac{\tau_i}{2\pi}\right)^2\right\}\right],$$

and

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

denotes summation over the 2^N discrete states of strand separation. Performing a matrix version of completing the squares evaluates $Q(n)$ as

$$Q(n) = \left(\left\{\frac{2\pi}{\beta C}\right\}^n \frac{4\pi^2 C}{4\pi^2 C + Kn}\right)^{1/2} \exp\left[\frac{-2\pi^2\beta CK}{4\pi^2 C + Kn}\left(\alpha + \frac{n}{A}\right)^2\right].$$

Calculation of the SIDD properties of DNA topoisomers requires the evaluation of several quantities. First, to find the transition profile, we calculate the equilibrium probability of denaturation of each base pair, which for base pair j is the ensemble average value $\bar{n}_j$ of the binary variable n_j. This is given by

$$p(j) = \bar{n}_j = \frac{Z_j}{Z}, \quad (5)$$

where

$$Z_j = \sum_{\mathcal{S}} \left\{ n_j Q(n) \exp\left(-\beta \sum_{j=1}^{N} \left\{ (a+b_j)n_j - a n_j n_{j+1} \right\} \right) \right\}. \tag{6}$$

Next, we evaluate the SIDD profile, which involves calculating the destabilization energies $G(j)$, $1 \leq j \leq N$, of each base pair in the domain. For the base pair at position j, $G(j)$ is defined as

$$G(j) = \bar{G}(j) - \bar{G}. \tag{7}$$

Here, $\bar{G}$ is the ensemble average value of the free energy of the system, and $\bar{G}(j)$ is the average free energy of those states in which base pair j is open. Thus, $G(j)$ may be viewed as the increment of free energy required to assure that base pair j remains always open. $\bar{G}$ is given by

$$\bar{G} = \frac{Z_G}{Z}, \tag{8}$$

where

$$Z_G = \sum_{\mathcal{S}} \left\{ Q_G(n) \exp\left(-\beta \sum_{j=1}^{N} \left\{ (a+b_j)n_j - a n_j n_{j+1} \right\} \right) \right\}, \tag{9}$$

with

$$Q_G(n) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} d\tau_i \; G \exp\left[-\beta \left\{ \sum_{i=1}^{n} \frac{C\tau_i^2}{2} + \frac{K}{2} \left(\alpha + \frac{n}{A} - \sum_{i=1}^{n} \frac{\tau_i}{2\pi} \right)^2 \right\} \right],$$

and the expression for G is given by Equation (3) above; $\bar{G}(j)$ is given by

$$\bar{G}(j) = \frac{Z_{G,j}}{Z_j}, \tag{10}$$

where

$$Z_{G,j} = \sum_{\mathcal{S}} \left\{ Q_{G,n_j} \exp\left(-\beta \sum_{k=1}^{N} \left\{ (a+b_k)n_k - a n_k n_{k+1} \right\} \right) \right\}, \tag{11}$$

with

$$Q_{G,n_j}(n) = \prod_{k=1}^{n} \int_{-\infty}^{\infty} d\tau_k \; G n_j \exp\left[-\beta \left\{ \sum_{k=1}^{n} \frac{C\tau_k^2}{2} + \frac{K}{2} \left(\alpha + \frac{n}{A} - \sum_{k=1}^{n} \frac{\tau_k}{2\pi} \right)^2 \right\} \right].$$

We note that $\bar{G}(j)$ is averaged only over those states where base pair j is open, *not* over all states.

The algorithm presented below to implement this analysis uses accumulators that are periodically updated to construct the partial sums for the partition function and the other quantities needed in these calculations. The accumulator for the partition function Z is denoted $\mathcal{A}(Z)$, that for Z_G is denoted $\mathcal{A}(G)$, and similarly for the others.

Other ensemble averages of interest may be calculated either directly by analogous methods or from the quantities already found. The average number $\bar{n}$ of open base pairs is given by

$$\bar{n} = \sum_{i=1}^{N} \bar{n}_i = \sum_{i=1}^{N} p(i).$$

The number r of open regions in a specific state of denaturation of a circular topoisomer is

$$r = \sum_{i=1}^{N} n_i(1 - n_{i+1}), n < N,$$

and $r = 1$ when $n = N$, the case where all base pairs are denatured. (Here we use the circularity condition that $n_{N+i} = n_i$.) Then, the ensemble average number of open regions is given by

$$\bar{r} = \frac{Z_r}{Z}, \tag{12}$$

where

$$Z_r = \sum_{\mathcal{S}} \left\{ r\,Q(n) \exp\left(-\beta \sum_{j=1}^{N} \left\{ (a + b_j)n_j - a n_j n_{j+1} \right\} \right) \right\}.$$

Evaluation of this quantity requires determination of one more sum, which is

$$Z_{n_i,n_{i+1}} = \sum_{\mathcal{S}} \left\{ n_i n_{i+1} Q(n) \exp\left(-\beta \sum_{j=1}^{N} \left\{ (a + b_j)n_j - a n_j n_{j+1} \right\} \right) \right\}.$$

The ensemble average twist $\bar{\tau}_i$ of an open base pair may be calculated using analogous procedures, with the continuous quantity τ_i to be averaged appearing inside the integrand.

Evaluation of the statistical mechanical quantities

Three techniques have been developed to evaluate the above quantities. First, we have previously presented a formally exact method that has quadratic computational complexity for calculating the transition probability profile and requires linear memory (Fye and Benham, 1999). Because the imposed superhelicity couples the behavior of each base pair to that of every other base pair, SIDD is inherently a quadratic process. So the computational complexity and memory requirements of this algorithm are both optimal. Unfortunately, this method suffers from a severe catastrophic cancellation problem, which requires calculations on sequences of kilobase lengths to be implemented in arithmetic of arbitrary precision. This slows execution time by at least two orders of magnitude, making the exact method impractical for all but the shortest sequences. In practice, its primary use is to calibrate the approximate method, as described below.

We also have developed a Monte Carlo method which does not calculate the equilibrium distribution, but rather samples it (Sun *et al.*, 1995). The average frequency with which individual states are selected decreases exponentially as their free energies increase, so a sufficiently long and unbiased sampling will converge to the equilibrium distribution. However, in practice, the rate of this convergence is extremely slow. Very long run times are needed to evaluate the frequencies of occurrence of improbable states. So this method cannot accurately evaluate the behaviors of sites experiencing subthreshold destabilization, i.e., where the energy needed to open the DNA duplex is diminished by an amount that is not sufficient for local denaturation. Subthreshold destabilization is biologically important because it can potentiate other interactions. So an informative method must accurately calculate sites and extents of destabilization for these locations also. Because the Monte Carlo method can yield only reliable estimates of the lowest energy states in reasonable run times, it also is unsuitable for informative large-scale calculations.

The most practical technique for performing the above calculations is an approximate method. Here, we develop a generalization of the original technique (Benham, 1990, 1992), which allowed only copolymeric energetics—one value b_{AT} for AT pairs and another b_{GC} for GC pairs. The present, more general method allows arbitrary assignments of the denaturation free energies b_i. The algorithmic implementation of this generalized method also has been carefully optimized for efficiency.

The approximate statistical mechanical method

The strategy of this calculation is as follows. First, the state of absolute lowest free energy is found, and its energy G_{min} is noted. Then, an energy threshold θ is specified, and the set $\mathcal{S}_\theta$ of all states s is found whose free energies $G(s)$ exceed G_{min} by no more than θ:

$$\mathcal{S}_\theta = \{\forall\, s \in \mathcal{S} \mid G(s) - G_{min} \leq \theta\}\,.$$

Approximate ensemble average (i.e., equilibrium) values of the partition function and of all quantities of interest are calculated using this subset of states. These calculations use the appropriate equations presented above (Equations (4)–(6), (8)–(11), (12)), with the summations limited to the set $\mathcal{S}_\theta$ of states satisfying the threshold condition. For example, the approximate partition function $\hat{Z}$ is evaluated as

$$\hat{Z} = \sum_{s \in \mathcal{S}_\theta} \exp(-G(s)/RT).$$

Parameters that can be evaluated in this way include the transition and stress-induced duplex destabilization (SIDD) profiles, ensemble average values of the residual superhelicity, the torsional deformation of the strand separated regions, the number of separated base pairs, and the number of runs of transition.

Although *individual* high energy states (i.e., those whose energies exceed the threshold) are exponentially less populated than are low energy states, because they are so numerous, their *cumulative* contribution to the equilibrium still may be significant. So the optimal use of this strategy requires a careful evaluation of how the aggregate influence of the excluded high energy states and the number of included states both scale with θ. We use the exact method for this purpose (Fye and Benham, 1999).

The number of states that are included in this analysis increases approximately exponentially with the threshold θ. Under physiologically reasonable assumed conditions, high accuracy is achieved using moderate values of θ, for which calculations execute efficiently. Calculations on many sequences under the conditions used in Kowalski's nuclease digestion procedure for finding stress-destabilized sites (Kowalski *et al.*, 1988) and at a mid-physiological superhelix density of $\sigma = \alpha/Lk_o = -0.06$ achieve four to five significant figures of accuracy in all parameters using a threshold of $\theta = 12$ kcal/mol. Calculations using these values performed on sequences of length $N \approx 5{,}000$ bp commonly will have somewhere between 10^6 and 10^9 states that satisfy this threshold, a number that is small enough to execute efficiently. This makes the approximate method the technique of choice under most circumstances.

We note that the algorithmic implementation of this strategy commonly includes more states than those satisfying the above inequality, which correspondingly improves the accuracy of the approximation. The reasons for this are described below.

This strategy may not be suitable for calculations that assume certain extreme conditions. In particular, at temperatures exceeding the melting temperature of duplex DNA, such as are experienced by thermophilic organisms, this method may become intractable. Under these conditions, the low energy states can be highly denatured, so their number could be excessive. The exact method is the only feasible strategy under such circumstances.

Algorithmic implementation of the approximate method

Preprocessing of the DNA sequence information. The function $E(j)$, $1 \leq j \leq N$, is defined as the sum of the opening energies of all the base pairs from 1 to j inclusive:

$$E(j) = \sum_{i=1}^{j} b_i = E(j-1) + b_j.$$

We specify that $E(0) = 0$ and extend to accommodate the wrap-around condition imposed by circularity as $E(j+N) = E(j) + E(N)$, $1 \leq j \leq N$.

Next, we construct an array containing the total opening energy of every window in the molecule of length l, $1 \leq l \leq l_{max}$. Here, l_{max} is chosen to be substantially larger than the largest number of open base pairs in any state satisfying the threshold condition. In practice, $l_{max} = 250$ bp is adequate for

analyzing $N = 5{,}000$ bp sequences at physiological superhelicities and environmental conditions, assuming a threshold $\theta = 12$ kcal/mol. Then, the window of length l starting at position i has total opening energy

$$X_l(i) = E(i + l - 1) - E(i - 1) + a.$$

This includes both the separation energies b_j of all base pairs in the run and the run nucleation energy a. For each window size l, these entries are sorted by energy, lowest to highest. From this, we construct an array $Y(k, l)$ which contains two entries at each position. The first entry $Y_1(k, l)$ is the opening energy of the kth-lowest energy window of width l, and the second entry $Y_2(k, l)$ is the position of the first base pair in that window.

Finding the lowest energy state. The free energy associated to a specific state (Equation (3)) contains both a discrete component and a continuous component. The discrete contribution G_d is the free energy required to denature the specific base pairs that are separated in that state. (See Equation (2).) For each specific state of denaturation (i.e., each specification of the values of the N binary variables n_j), according to the conservation equation, Equation (1), there is a continuum of ways in which the residual deformation may be partitioned between twisting τ_j of the open base pairs and the residual superhelicity. The free energy associated to this continuous component is

$$G_{\mathcal{T}} + G_r = \frac{C}{2} \sum_{k=1}^{n} \tau_k^2 + \frac{K}{2} \left(\alpha + \frac{n}{A} - \sum_{k=1}^{n} \frac{\tau_k}{2\pi} \right)^2. \tag{13}$$

(Summations are over the n open base pairs.) This expression describes a symmetric paraboloid whose domain is the n-dimensional τ_k-space, $1 \le k \le n$. It has a unique absolute minimum value at the point where all its first partial derivatives vanish. A simple calculation shows that at this point all τ_k's have the same value $\tau_k = \tau_{min}$, $1 \le k \le n$, which is

$$\tau_{min} = \frac{2\pi K \left(\alpha + \dfrac{n}{A} \right)}{4\pi^2 C + Kn}.$$

Placing this value into Equation (13) above and simplifying shows the minimum value of the continuous portion of the free energy expression to be

$$G_{min,\tau_k}(n) = \frac{2\pi^2 CK}{4\pi^2 C + Kn} \left(\alpha + \frac{n}{A} \right)^2,$$

which for a fixed value of the superhelicity α is a function only of the total number n of open base pairs in the state.

Next, we consider the discrete components of the states. We denote by $G_{min}(n, r)$ the free energy of the lowest energy state in which n open base pairs occur in r distinct runs. (A run is a set of contiguous open base pairs.) This will include the energy of opening the base pairs involved, if any, plus the minimal stress energy $G_{min,\tau_k}(n)$ found above. For each value of r and n, we use the array Y found above to determine the energy $G_{min,o}(n, r)$ of opening of the lowest energy (n, r)-state. This is done progressively, first for increasing values of n in the $r = 1$ run states, then in the two-run states, and then in the three-run states. In each case, we need consider only states whose total number of open base pairs does not exceed n_{max}, which in the implementation below is $n_{max} = 250$ bp. Once this calculation has been completed for r-run states, we denote the minimum energy found to that point by $G_{min}(r)$, with $G_{min}(0) = K\alpha^2/2$. In practice, we need consider only states with $r \le 3$, as, for reasons described below, only these are significantly populated at equilibrium under physiological conditions and stress levels.

The analysis of one-run states. Recall that the jth column of the array Y contains the sorted opening energies of the windows of length j, together with their locations. So the lowest energy state with n open base pairs in a single run is given by the first entry in column n, and $G_{min,o}(n, 1) = Y_1(1, n)$. The total energy is found by adding the minimum stress energy to this quantity,

$$G_{min}(n, 1) = G_{min,o}(n, 1) + G_{min,\tau_k}(n),$$

and the absolute minimum energy state found to this point is the minimum of these values:

$$G_{min}(1) = \min\{G_{min}(0), G_{min}(n, 1),\ 1 \le n \le n_{max}\}.$$

We note that $G_{min}(1)$ is an upper bound for the energy of the absolute minimum energy state, which may have more than one run.

Next, we collect all the information about one-run states that will be needed to calculate their contributions to the partition function and to the other sums used to calculate all ensemble averages of interest. Recall that in this approximate strategy we explicitly include all states whose free energy exceeds the minimum value by no more than a specified threshold amount θ. (As will be seen below, we also may include some states which do not satisfy the threshold condition.) To find these states, for each run length n, we use $Y_1(k, n)$, which gives the opening energy of the k-th lowest energy run of length n. We add to this the *minimum* stress energy $G_{min,\tau_k}(n)$ and check that the resulting total energy satisfies the threshold condition, which we calculate using the minimum energy $G_{min}(1)$ found to this point:

$$Y_1(k, n) + G_{min,\tau_k}(n) \le G_{min}(1) + \theta. \tag{14}$$

For each value k satisfying the threshold condition, we calculate its associated Boltzmann factor, which is the contribution $Z(n, 1, k)$ to the partition function from all states with this specific single run of denatured base pairs:

$$Z(n, 1, k) = Q(n) \exp(-\beta Y_1(k, n)). \tag{15}$$

This explicitly includes all states of residual stress distribution in the factor $Q(n)$, some of which will in fact exceed the threshold. We update the accumulator $\mathcal{A}(Z)$ that is summing these contributions by adding $Z(n, 1, k)$ to it. We also determine the contributions these states make to the sums needed to calculate the other quantities of interest and update their accumulators.

Next, we move down the array to the $k + 1$-st entry and test whether it satisfies the threshold condition using Equation (14). If it does, we calculate its Boltzmann factor and related quantities and update the accumulators as above. As the columns of the array Y have each been sorted by increasing energy, the energies of successive states encountered in this way will be monotonically increasing. So we can stop as soon as the threshold condition of Equation (14) fails. At this point, we increase the run length by 1, which uses the $n + 1$st column in the Y array, and repeat this process. Proceeding in this way, we consider successively longer one-run states, up to the maximum n_{max}. If, for a given length n, the first entry in column n of the Y array exceeds the threshold, then no $(n, 1)$-states will satisfy the condition, so we immediately move on.

Use of this strategy has two implications. First, the inequality in Equation (14) above finds those states whose energies satisfy the threshold condition when compared to the minimum energy state found to this point, *not* to the absolute minimum energy state. If it is not the absolute minimum free energy, then this strategy will find more states than if the true minimum were used. And second, this inequality includes only the lowest energy stressed state, not the distribution over all possible states of stress. As all states of stress are included in the expressions to be added to the accumulators (viz. Equation (15) above), this again includes states whose energies exceed the threshold. Inclusion of states whose energies are above the threshold yields a corresponding increase in the accuracy of the results.

The analysis of two-run states. We treat the two run (i.e., $(n, 2)$-) states in a similar manner. The analysis of the stress energy is identical to that presented above, as both $G_{min,\tau_k}(n)$ and $Q(n)$ do not depend on the number r of runs in which the n open base pairs occur. However, the analysis of the discrete states is more complicated. First, we must specify the lengths of the two runs, which involves partitioning $n = n_1 + n_2,\ 1 \le n_1 \le n_2 \le n_{max}$. There are $[n/2]$ ways to do this, each of which must be considered separately. (Square brackets here denote the greatest integer function.) Recall that the top entries in the n_1 and n_2 columns of the Y array give the lowest energy runs of those lengths, and suppose that the energy of the absolute lowest energy state found to this point is $G_m \le G_{min}(1)$. If the total energy of the state they determine does not satisfy the threshold condition

$$Y_1(1, n_1) + Y_1(1, n_2) + G_{min,\tau_k}(n) \le G_m + \theta, \tag{16}$$

then we know there are no $(n, 2)$-states with this (n_1, n_2) partitioning that satisfy the threshold. In that case, we proceed directly to the next partitioning. Once all the partitionings have been examined (i.e., $1 \leq n_1 \leq [n/2]$), we proceed to the next value of n and continue in this way until the maximum n_{max} is reached.

Suppose the threshold condition expressed by Equation (16) is satisfied for the top entries in the two columns. Then, the runs involved define a true two-run state only if they do not overlap or abut. So we must test every such pair of runs that satisfy the two-state threshold condition to determine whether they are in fact distinct. If they are distinct, we calculate the contributions to the accumulators of the states having these open runs as

$$Z(n_1, 1, n_2, 1) = Q(n) \exp\left(-\beta(Y_1(1, n_1) + Y_1(1, n_2))\right)$$

and update accordingly. Then, we move down the n_2 column, testing n_2-states of increasing energy, keeping the n_1 entry at the top position. Each state that has distinct runs is accumulated in this way, until the position is reached where the energy of the minimum energy state with these runs exceeds the threshold. This is determined from Equation (16) using the appropriate entries Y_1. When this happens, the n_2 entry is put back to the top of its column, we move one position down the n_1 column, and repeat this process. This procedure is continued until the current entry in the n_1 column, together with the top entry in the n_2 column, exceeds the threshold. If any time true state is found whose energy is lower than G_m, we update G_m accordingly. Then, we move to the next subdivision, increasing n_1 by unity and decreasing n_2 by unity, until all possible subdivisions have been examined. Then, we move to the next value of n. When all values $2 \leq n \leq n_{max}$ have been examined, the two-run state analysis is complete. The state of lowest free energy found to this point is $G_{min}(2)$.

The analysis of three-run states. We treat the three-run states in a strictly analogous manner. Here, the only differences are 1) that there are many more ways to partition $n \geq 3$ open base pairs into three-runs: $n = n_1 + n_2 + n_3,\ 1 \leq n_1 \leq n_2 \leq n_3$, and 2) that in any true three-run state no pair of runs can overlap or abut, so there are three pairwise tests of distinctness to perform in each case.

Extensive sample calculations performed on many different DNA sequences of lengths $N \leq 10,000$ bp assuming physiological conditions and stress levels have found that even with the extreme threshold of $\theta = 20$ kcal/mol, states with more than $n_{max} = 250$ total open base pairs or more than three runs of opening do not occur. This is true because the nucleation energy $a = 10.2$ kcal/mol greatly exceeds the energy b_j of opening a base pair. Dividing n into more runs increases the number of nucleation events that must occur, which are energetically very expensive. To compensate, the runs involved must have a substantially decreased total separation energy. As the low energy one-run states have low separation energy to start with, it often is difficult to decrease this substantially by dividing the single run into two or more runs, perhaps placed elsewhere. Decreases sufficient to fully compensate for the nucleation energies required to open the additional runs usually are not possible (except perhaps in extreme circumstances), so states with more runs commonly have higher total energies. As the number of runs increases, this effect becomes more severe. In practice, only states with $r \leq 3$ occur for $\theta \leq 20$ kcal/mol under normal conditions.

In the vast majority of case examined to date, the lowest energy state was either untransformed or had a single run of open base pairs. In a few cases involving extreme negative superhelicity (viz. $\sigma = \alpha/Lk_o \approx -0.1$), a two-run state had minimal energy, but in no case was a three-run state found to be minimal.

The complexity of the approximate method increases rapidly with the number r of runs that must be considered. It is only because r_{max} is small that this strategy is computationally tractable. Thus, it may not be the optimally efficient approach to analyze superhelical strand separation under thermophilic conditions, where the temperature exceeds the melting temperature of DNA (Benham, 1996b).

III. THE MULTI-WINDOW METHOD OF SIDD ANALYSIS

Long DNA sequences, including complete chromosomes, are analyzed by partitioning them into windows and analyzing each window separately. All windows are chosen to have the same length N. Successive windows are offset by a distance d_o, with $d_o|N$, so each internal base pair appears in $w = N/d_o$ windows. Here, we use the above-described approximate method to analyze each window, although other techniques

also could be used. Successive windows are analyzed until the entire sequence has been traversed. The overall size of the total calculation scales linearly with the number of windows involved. The final values of the opening probability $p(x)$, the destabilization energy $G(x)$, and any other parameters of interest for the base pair at position x are calculated as weighted averages of the values computed for the windows containing that base pair:

$$p(x) = \sum_{i=1}^{w} W_i p_i(x), \tag{17}$$

and

$$G(x) = \sum_{i=1}^{w} W_i G_i(x). \tag{18}$$

To apply this windowing procedure, we must specify the window size N, the offset distance d_o, and the weighting function W_i, and we also must treat the sequence ends differently for linear and circular chromosomes. Below we examine the implications of each of these choices.

It is not our intention at present to explicitly mimic *in vivo* conditions. Indeed, one anticipates these will vary in complex ways according to the precise manner in which stresses are imposed, how the DNA is constrained, binding events, and many other specific factors. These calculations are presently intended to illuminate a relatively simple physical chemical attribute of the DNA duplex—its propensity to become locally destabilized in response to the torsional stresses that occur in living systems. Although the assumptions implicit in this approach are much simpler than is the *in vivo* situation, their results have illuminated attributes of the DNA that have been implicated in a variety of important regulatory processes, as described in the Introduction above. Accordingly, we design this windowing procedure to be maximally effective at illuminating sites whose susceptibilities to stress-induced duplex destabilization (SIDD) may be involved in their mechanisms of activity. This is the perspective from which we examine the implications of the various choices made in designing a specific windowing method.

First, consider the influence of the window size N. Recall that the behaviors of all base pairs within a single window are directly coupled together by the stresses they experience. Under physiological conditions the number of open base pairs commonly does not exceed 3% of the total. So if the window size is small, denatured states will contain relatively few open base pairs. The difference in opening energy between alternative short sites is correspondingly small, so this competition is between many fractionally destabilized sites. (A similar phenomenon also occurs at the onset of transition in longer sequences, as there also the expected number of open base pairs commonly is small.) Conversely, the behavior of a very long window will be dominated by states having proportionately longer runs of opening. The energy of opening of long runs in average genomic sequences will commonly vary greatly from one potential site to another, so the open states are dominated by the few most easily destabilized sites. It has been our experience that an intermediate window size is most informative regarding the relative propensities of sites to open. Accordingly, here we select $N = 5{,}000$ kb as the most suitable window size.

Next, consider the offset distance d_o, noting that each base pair (except perhaps those near the sequence ends) will occur within N/d_o windows. The transition behaviors of two base pairs are *directly* coupled by imposed stresses when they both occur within the same window. So the total strength of this coupling will be determined by the number of windows they share, and the weight function assigned to the windows. (We note, however, that if $d_o \leq N/2$, a global, indirect coupling exists, even for sites separated by more than N base pairs so they never appear in the same window. This is the result of a chain of couplings from one base pair to its near neighbors, then from them to their neighbors and so on, all the way to the other base pair.) There are several possible choices for the offset distance, each of which makes implicit assumptions regarding this coupling. We consider several alternatives in turn.

First, choosing $d_o = N$ selects successive N base pair windows with no overlap. While this is computationally the fastest strategy, it has unacceptable implications regarding the coupling induced by the stress. Specifically, it ascribes perfect coupling to all the base pairs within a window and complete decoupling between windows. So a base pair near the edge of a window (say at position 4,900 in a 5,000 bp window) will be perfectly coupled to base pairs positioned more than 4,500 bp to its left, but totally uncoupled from

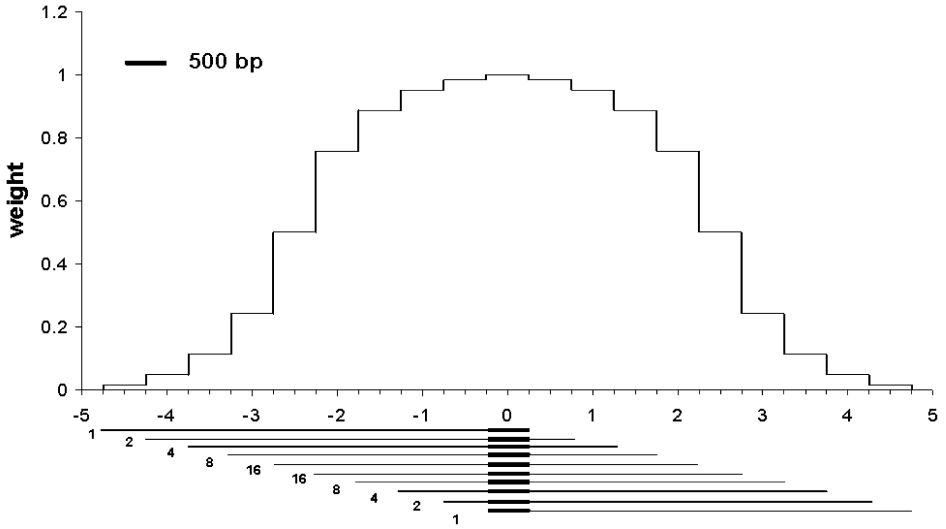

FIG. 3. The windowing procedure analyzes successive 5,000 bp windows, each offset by 500 bp so an internal base pair occurs in exactly 10 windows as shown by the bars at the bottom of this figure. The relative weights of individual windows increase exponentially as the base pair proceeds from peripheral to central, so the total weight function for all regions that contribute to the final values for a given base pair vary with the distance from that base pair as shown in the top part of the figure.

base pairs located 200 bp to its right. This can eliminate interactions between nearby sites that may in fact be strongly coupled and overstate interactions between more remote sites which might influence each other only weakly. Moreover, it does this in a highly asymmetrical manner for the base pairs near the edges of any window. To illustrate the problem, consider two chromosomes which differ only by a 2 kb indel. This difference offsets the edge positions of all downstream windows in one chromosome relative to the other. This in turn changes the couplings, and hence potentially the calculated destabilization properties, at all downstream positions—even those many megabases away. Further, which side of the indel is regarded as being downstream depends on the orientation of the chromosome. So the results using the $d_o = N$ offset in this example could vary in unacceptable ways with attributes of the calculation strategy. For these reasons, the nonoverlapping choice is not satisfactory.

Next, consider the choice $d_o = N/2$. Here, every base pair appears in two windows, each of which by symmetry must be weighted equally, so the weight function is $W_i = 0.5$. The overlap between these windows is $N - d_o = 2{,}500$ bp, so each base pair will be coupled at some strength to all others within a region spanning 7,500 bp. It will be fully coupled in the central 2,500 bp where it is located and coupled with strength $1/2$ to the peripheral 2,500 bps on either side. Here again, considerable asymmetry may occur, depending on where within the central 2,500 bp the base pair in question is located. If it is near the edge, then it can be fully coupled to base pairs more than 2 kb on one side, but coupled at strength 1/2 to much closer sites in the other direction. However, it will be coupled at some strength to all the base pairs for at least 2.5 kbp on each side, so the maximum severity of the possible asymmetry is substantially reduced from that occurring in the nonoverlapping case.

As a third alternative, suppose $d_o = N/10$, so each base pair appears within ten windows that span a total of 9,500 bp. This case is illustrated in Fig. 3. We assign larger weights to windows in which the base pair is central and smaller weights to ones in which it is peripheral. Here, we assign exponential relative weights of 1, 2, 4, 8, 16, 16, 8, 4, 2, 1 as the windows proceed from left to right across the base pair of interest. Because the weights must add to unity, their values W_i are found by dividing these numbers by their sum, which is 62. At the bottom of Fig. 3 are shown the ten windows that all contain the highlighted 500 bp segment. These 500 bps are present in all ten windows, so the sum of all the weights of the windows in which they occur is unity—meaning that they are fully coupled together. The 500 bp region to its immediate left appears in every window but the rightmost one, so the sum of the weights of the windows in which it occurs is 61/62, indicative of a strong coupling between the highlighted segment

and the central region. A symmetrically positioned and equally weighted 500 bp region is found to the immediate right of the central site. As one proceeds further from the central region, there are symmetrically placed 500 bp regions which co-occur with the base pair of interest in successively fewer windows, and have correspondingly smaller total weights. There will be a total of 9,500 base pairs that co-occur with the pair in question in at least one window. The total weights of their couplings to the base pairs in the central, highlighted segment are plotted at the top of Fig. 3. These choices have the effect of coupling each base pair strongly to its nearer neighbors, with the strength of the coupling falling off approximately sigmoidally with distance as shown.

In developing the windowing method, we have tried several different weighting schemes. Arrangements in which the W_i's increase monotonically and symmetrically as the base pair in question moves from peripheral to central are the only physically reasonable choices. However, extensive sample calculations showed little practical difference if the individual weights increase exponentially as above, or quadratically, or even linearly.

There are significant tradeoffs involved in increasing the number of windows. Although this strategy decreases the size of the maximal asymmetries that can occur, it increases the number of windows needed to span the sequence, and hence the total execution time. Our sample calculations suggest that the optimal choice lies between five-fold and ten-fold coverage. In the algorithm implementing the windowing procedure, we choose default values $N = 5{,}000$ bp and $d_o = 500$ bp, which gives ten-fold coverage.

Finally, consider how to treat chromosome ends. Circular DNA sequences, as occur with certain viral and prokaryotic chromosomes, require a simple wrap-around condition. Here, the first 5,000 bp is duplicated and placed at the end of the sequence, and the last 5,000 bp is duplicated and placed at its beginning. In this way, every base pair is analyzed in its natural sequence context, and every pair occurs within the same number of windows.

A different strategy is required for linear chromosomes or parts of chromosomes, where two alternatives suggest themselves. First, one can proceed by successive windows from the start to the end of the actual sequence. This analyzes each position in its true context, but it treats end-proximal regions differently from end-distal ones. For example, the 500 bp region at the start of the sequence will appear in exactly one window, the next 500 bp in two windows, etc. Only those base pairs after position 4,500 appear in a full set of ten windows. So the first (and last) 4,500 bp are perforce treated differently than the others. Because the weights still must add to unity, the weighting scheme used must be different in these end-proximal regions from those at more internal sites. The end-most 500 bp appear in only one window, so the base pairs in that region are fully coupled to every base pair in that 5,000 bp window, then totally uncoupled from all more remote base pairs. Similarly, the base pairs between positions 501 and 1,000 appear in exactly two asymmetric windows, whose total weights must again add to unity, and so on. This approach has the advantage of analyzing each base pair in its actual sequence context, but the disadvantage of treating the 4,500 bp on each end of the sequence differently from those of the more internal regions.

A second strategy is to increment the actual sequence by placing 4,500 bp of additional sequence both before its start and after its end. The base sequence of the incremental region can be chosen in various ways, one option being random selection with base frequencies equal to those of the actual sequence. In this case, every base pair of the actual sequence occurs in ten windows, so a uniform weighting scheme can be used throughout. This means the couplings between end-proximal regions and their true neighbors have the same weights as those used for internal portions of the sequence. However, here the aggregate behaviors of the end-proximal regions are determined in part by competitions with fictitious sequences that have been artificially added.

In practice, either of these two strategies may be adopted. However, in both cases it must be recognized that the telomeric 4.5 kb regions are treated differently from the more internal sequences. So the results computed for those regions must be evaluated accordingly. The default strategy used in the implemented algorithm is the second one, augmenting the sequence on each end with 4,500 bp of random sequence DNA.

Computational implementation of the windowing procedure

The windowing procedure described above has been implemented in C++ with the following default values: $N = 5{,}000$ bp, $d_o = 500$ bp, and $\theta = 12$ kcal/mol. The opening energies may be chosen to be either copolymeric (one value for every AT base pair and a different value for every GC base pair), or specified individually. The precision of the floating point arithmetic needed for implementation depends on

the choice made of the threshold θ. As described above, $\theta = 12$ kcal/mol has been selected as the default because it gives good accuracy with reasonable per-window execution times. All states whose free energies exceed the minimum by less than this amount will be explicitly included in the calculation. The equilibrium frequency of occupancy of a state at this threshold is $f = \exp(-12/RT)$ times that of the lowest energy state. At room temperature, $RT \approx 0.6$ kcal/mol, so $f \approx 2 \times 10^{-9}$. Contributions to accumulators that are of this size will be lost in the underflow unless the summations involved use at least 64-bit floating point arithmetic. So double precision is required on a 32-bit processor using this threshold, but only single precision is needed on a fully 64-bit processor.

The windowing algorithm is well suited for parallel implementation on a distributed memory architecture machine. The sequence can be divided into segments by the master processor, and the calculations for individual segments are performed indsependently on separate processors. If the segments are individual windows, then the results are sent back to a master, which amalgamates them by performing the weighting. In practice, it is more efficient to analyze longer segments, performing the weighting on the individual processors and collecting the final results on the master unit. The implementation used in our calculations uses 100 kbp segments with 5 kbp overlaps on each end. Segments are assigned to processors successively as they complete their previous assignments. This coarse-grained implementation is naturally load balanced, as all processsors are fully occupied until the last few segments are assigned.

Because of the complexity of the algorithm and the large amount of resources needed to implement it efficiently for long sequences, we do not at present contemplate routinely providing source code or executables to users for remote implementation. However, a website is currently under development that will give access to the results of SIDD calculations on complete chromosomes from a variety of organisms. Users will be able to specify the organism, chromosome number, and starting and ending positions and receive graphs or tables of SIDD parameters (i.e., $p(x)$ or $G(x)$) at single-base resolution. This website will be found at www.genomecenter.ucdavis.edu/benham/.

IV. SAMPLE CALCULATIONS

Parameter values

In our calculations, we use the experimentally determined values of all parameters that are appropriate for the base sequence of the molecule under the assumed environmental conditions. Because these free energies have been experimentally evaluated, this model has no free parameters. The default option is to use the conditions of Kowalski's mung bean endonuclease digestion procedure (Kowalski *et al.*, 1988) which is the most accurate experimental method currently available for detecting unpaired bases in superhelical DNA. These are T = 37C, $[Na^+] = 0.01$ M, and pH = 7.0. However, any conditions can be asssumed for which accurate free energy values of all parameters either have been measured or can be interpolated from measurements.

The denaturation free energies b_j have been experimentally shown to depend upon base identity, near neighbors, ionic conditions, temperature, and the presence of imperfections such as adducts, lesions, mispaired or unpaired bases, or abasic sites. Both the entropy and enthalpy of denaturation have been measured to high accuracy for all ten types of neighbor base pairs at several ionic strengths (Steger, 1994), which permits evaluation of b_j at any temperature. In our calculations, we select the values measured by Klump as our defaults. These do not assume constant entropies of opening, as do others, and they were evaluated at ionic strength 0.075 M. This is intermediate between 0.01 M, the strength used in the Kowalski nuclease digestion procedure, and the *in vivo* value of 0.15M. Extrapolation to either of these important ionic strengths is straightforward (Schildkraut and Lifson, 1968) and for the Klump energetics involves relatively small changes.

Kowalski and coworkers have performed a careful experimental assessment of the dependence of the locations and extents of strand separation in pBR322 DNA on the level of imposed superhelicity (Kowalski *et al.*, 1988). In a previous paper, we determined the values of the free energy parameters k, a and C that gave the best fit to these experimental results assuming copolymeric denaturation energetics (Benham, 1992). Using the more general method presented here, we can assign transition energetics in a base-specific manner. We have repeated this fitting procedure using the experimentally measured near neighbor-dependent enthalpies and entropies of Klump (Steger, 1994), corrected to $[Na^+] = 0.01$ M. We have fit to Kowalski's

experimental measurements at three superhelicities, $\Delta Lk = -26$, -28 and -31, respectively (Kowalski *et al.*, 1988). The values that give the best fit to the data are: $a = 10.16$ kcal/mol., $K = 2{,}200\ RT/N$ kcal/mol., and $C = 1.91$ kcal/mol. These values agree very well with other experimental measurements of these parameters.

Comparison of the near neighbor and copolymeric energetics

The exact method for analyzing superhelical denaturation provided the first opportunity to assess how the properties of the SIDD transition depend on the assumptions made concerning the denaturation energetics. The paper presenting that method compared the results of exact calculations using copolymeric energetics to those assuming near neighbor energetics (Fye and Benham, 1999). The results calculated under these two assumptions did not differ significantly. Only at the edges of denatured regions were small boundary effects discernible. This close accord is not surprising, as the copolymeric energetics implicitly included averages of the near neighbor effects.

We have performed calculations using the approximate method presented above to compare the results for copolymeric energetics with those for near neighbor energetics. As with the exact method, the probability profiles calculated using these two energetics are almost indistinguishable. The results for the pBR322 DNA sequence at $\Delta Lk = -26$ turns using threshold $\theta = 12$ kcal/mol as calculated using these two energy functions are shown in Fig. 4.

These calculations revealed three differences between the copolymeric and the near neighbor cases. First, the energy parameters that were found to provide the best fit with experiments were in both cases close to their values as measured by other procedures. But they were somewhat closer when near neighbor denaturation energetics were used, as described above. This suggests that the near neighbor energetics provide a slightly more precise depiction of the SIDD properties of sequences. Second, although the

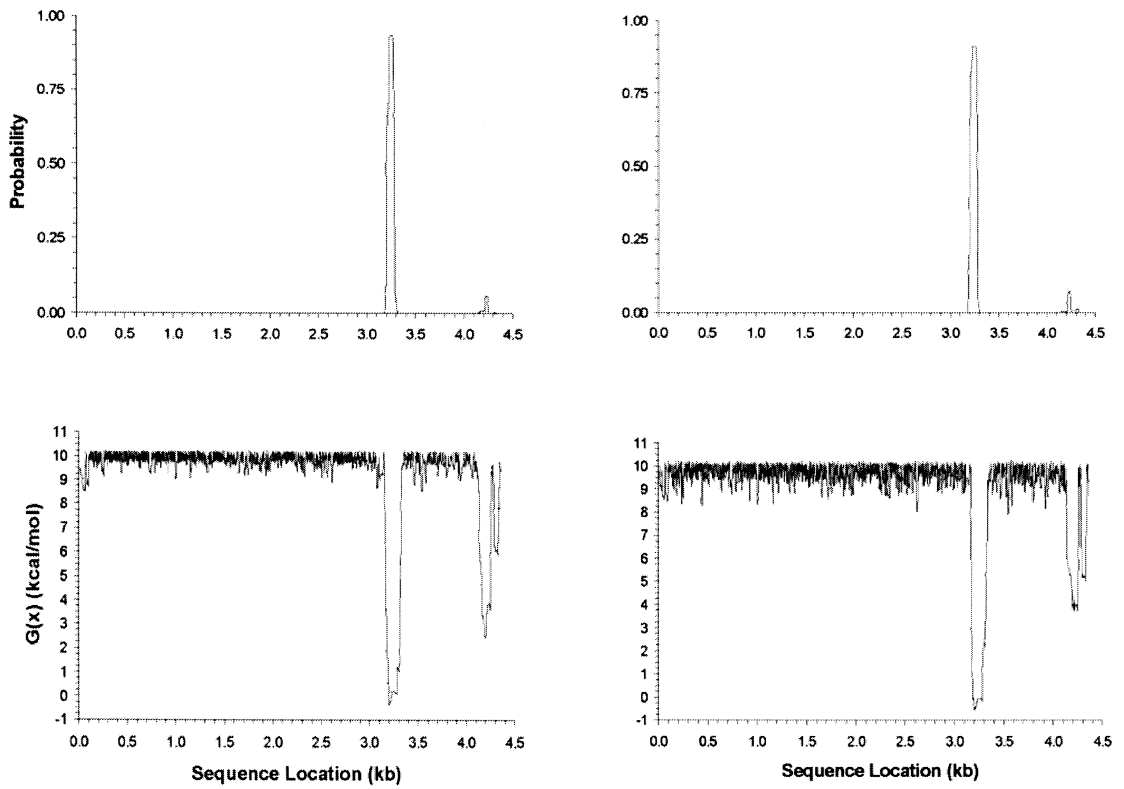

FIG. 4. The transition profiles (**above**) and the helix destabilization profiles (**below**) are shown for pBR322 DNA at $\alpha = -26$ turns as calculated using copolymeric energetics (**left**), and near neighbor energetics (**right**). One sees only a slight difference in the calculated destabilization energy $G(x)$ for the sites of intermediate stability at the extreme right.

denaturation probability profiles are almost indistinguishable in the two cases, the SIDD profiles do show minor differences. In particular, the profile calculated using near neighbor energetics shows more variation in the stable areas, and the destabilization at partially destabilized sites shows slight differences. Both of these effects would be expected from the use of more complex energetics, and both are consistent with earlier results using the exact method. The third difference is that more states are included in the near neighbor calculations than in the copolymeric ones. In our sample calculations, the ratio of these numbers of states varies from 2.24 at $\Delta Lk = -26$ turns to 6.01 at $\Delta Lk = -31$ turns. This difference also is an expected consequence of the more complex energetics. When these energetics are spread across ten different near neighbor types, there are many more ways to find low energy states than in the copolymeric case, where only two extreme values are used.

Analysis of the E. coli *genome*

As a first application of the windowing method, we have analyzed the complete *E. coli* K12 genomic DNA sequence, which is comprised of 4,639,221 base pairs in a circular configuration. The circular boundary condition was treated by duplicating the last 5,000 bp and placing them at the start of the sequence, then duplicating the first 5,000 bp and placing them at the end of the sequence. This enabled every window to be analyzed in its correct genomic context. This calculation involved 9,290 windows.

The complete analysis at superhelix density $\sigma = -0.055$ required 40.15 CPU hours to run on a dedicated 1 GHz Pentium III processor using the GNU C++ compiler at optimization level 2, with double precision floating point arithmetic. It found 81,101,140,512 states in total. The same calculation was executed at the rate of 1.1 Mbp/hour on a 28 node Apple X-Serve cluster. However, the execution time was found to depend significantly on the assumed stress level. When the calculations were performed at superhelix density $\sigma = -0.06$, they took approximately four times longer to complete.

This analysis finds that destabilization is confined to a small fraction of the genome, with the large majority of base pairs showing essentially no destabilization. A total of 282,759 bps, comprising only 6.095% of the entire genome, have $G(x) \leq 4.0$ indicative of substantial destabilization. Only 92,877 bp, or 2.002% of the genome, are strongly destabilized, having $G(x) \leq 1.0$. Conversely, more than 90% of the genome shows no significant destabilization. Moreover, the locations of SIDD sites showed statistically highly significant associations with specific classes of regultory regions, most notably promoters. For example, of the 190 intergenic regions known to contain experimentally documented promoters in the *E. coli* genome, 152 are destabilized below $G(x) \leq 4.0$. This is 18.6 standard deviations above the level of association that would occur if SIDD sites were randomly distributed relative to promoters. The analysis of these associations is ongoing; their results will be published in detail in future contributions.

V. DISCUSSION

This paper presents a computational method to evaluate the stress-induced duplex destabilization properties of long genomic sequences, including complete chromosomes. This method applies a windowing procedure in which individual windows are analyzed using a maximally flexible refinement of the approximate technique that enables the sequence-specific assignment of denaturation energies. Previous implementations of the approximate technique were confined to using copolymeric energetics. The sample calculations performed here and elsewhere suggest that results calculated using copolymeric energetics are not significantly different from those found using full near neighbor denaturation energetics. The latter choice provides a small but potentially informative improvement in assessing the SIDD properties of moderately destabilized regions.

However, the ability to assign denaturation energetics in a base pair-specific manner will allow a variety of potentially informative calculations that are not possible with copolymeric energetics. In particular, the effects of sequence modifications such as methylation, abasic sites, pyrimidine dimers, and other adducts or lesions can be assessed in all cases where their influences on the energetics of denaturation are known. These alterations could exert either regulatory or deleterious effects in part through modification of the SIDD behavior of the regions where they occur. In particular, their presence could alter or disrupt the normal SIDD activities involved in mechanisms of regulation, such as those described in the Introduction.

The ability to calculate the SIDD properties of complete chromosomes will enable the precise evaluation of how these properties correlate with the activities of all known types of DNA regulatory regions. This approach removes the possibility of selection bias arising from the analysis of small samples. It also could illuminate correlations between destabilized sites and regulatory regions. Such correlations have been repeatedly found in previous studies and often indicate that stress-induced destabilization plays a significant role in the governing regulatory mechanism. The availability of SIDD profiles for complete chromosomes and genomes will enable rigorous assessments to be made of the statistical significance of all such correlations between specific SIDD properties (perhaps together with other attributes) and the activities of every known type of regulatory region. Strong correlations, as have already been documented in specific cases, could then be used to develop accurate algorithms to computationally predict the locations of regulatory regions within previously uncharacterized genomic sequences.

In our first application of the windowing procedure, we analyzed the complete *E. coli* genome. A preliminary analysis of the results showed that promoters are associated with SIDD regions in a highly statistically significant manner. This suggests that SIDD attributes may prove useful in the computational prediction of promoter locations in prokaryotes, a problem that has proven surprisingly difficult to resolve using string-based methods alone. If there is a functional reason for this association between SIDD sites and promoters, as has been found in specific cases described in the Introduction, then there may be regulatory differences between promoters that are destabilized and those that are not. These issues, and many others arising from the analysis of this dataset, will be addressed in future work.

A partial analysis of several yeast genes shows that the most strongly destabilized sites in that eukaryote do not occur at promoters, but rather in the 3' terminal flanks of genes (Benham, 1996a). So *E. coli* and yeast show distinct patterns of association of SIDD sites with regulatory regions, each of which is statistically highly significant. These species- or kingdom-specific differences suggest that SIDD may play distinct roles, and perhaps also be regulated in different ways, in different organisms.

To analyse the complete human genome at the rate of 1.1 Mbp/hour achieved on the Apple X-Serve cluster would require 3,000 hours, approximately 18 compute-weeks. (We anticipate some speed-up from G4 processor-specific optimizations that have not been implemented yet.) At this rate, the analysis of all the sequenced prokaryotic genomes will require approximately two months, and analysis of the genomes of the sequenced eukaryotes will take more than one year. We intend to implement the SIDD windowing analysis program on a more high-throughput machine in the near future, so the results may be made available to the research community more quickly. Access to these results will be provided over the web at *www.genomecenter.ucdavis.edu/benham*. Once this site is opened for general access, users will be able to display the SIDD and probability profiles of any regions of interest within any chromosome that has been analyzed. They also will be able to download the $p(x)$ and $G(x)$ values for any region of interest so they can assess offline the SIDD properties of regulatory regions or of any other sites of interest. This website will initially contain the profile data for the complete *E. coli* chromosome, with other genomes to be added as their analyses are completed. The 16 chromosomes of the yeast *Saccaromyces cerevisciae* will be the first to be added. A second site also is being developed where users can submit shorter sequences $N < 10,000$ bp) for calculation using the approximate procedure described above. This site also will be accessible through the above web address.

The windowing strategy we have developed for analyzing long sequences regards the direct coupling between base pairs induced by stresses as diminishing with their separation distance, much as the apparent brightness of a light diminishes when it is viewed in a fog. It is reasonable to imagine that torsional stresses in DNA may behave in a similar manner. Stresses are unlikely to communicate the entire length of a chromosome, as both prokaryotic and eukaryotic chromosomes are partitioned into domains. However, even within a domain, one might expect that the strengths of stress-induced couplings could decrease with separation distance, although the scale of this effect is not known. So the assumptions implicit within these choices may accord roughly with *in vivo* conditions. However, if experiments show the specific locations of domain boundaries or other modulators of this coupling, these can be incorporated directly into the calculations.

In this paper, we do not explicitly consider what calculation parameters (window size, stress level, etc.) might best accord with biological reality. Indeed, although the *in vivo* situation is presently unknown, it is certain to be complex and may vary according to the phenomenon being investigated. There are many different ways that stresses are imposed on DNA, of which static superhelicity is only a paradigm example.

Dynamic waves of supercoiling are driven by the translocation of polymerase molecules. Distributions of stress can be significantly affected by many processes, including the arrangement of nucleosomes and other architectural molecules, the binding or release of regulatory proteins, the activities of nucleases, helicases, and single strand-specific binding proteins, the formation of adducts or lesions, and other events. These attributes may change in complex ways in different domains of a chromosome, with cell type, developmental stage, through the cell cycle, etc. For these reasons, the present calculations are not meant to reflect the full complexity of *in vivo* interactions, although that may become feasible later as more information about *in vivo* conditions becomes available. Instead, the method presented here should be regarded as a means to calculate a relatively simple physical chemical attribute, SIDD. Its intended use is to computationally search genomic DNA sequences for those sites whose activities may involve stress-induced duplex destabilization, as has been shown already for several specific classes of regulatory regions. Their results will enable rigorous assessments to be made of the statistical significance of correlations between SIDD properties and regulatory or other classes of sites. These could suggest experimental tests to determine the precise role that destabilization might play in specific mechanisms of regulation.

The results found to date have documented the role of SIDD in a wide variety of normal and pathological processes. A possible reason why SIDD appears to be implicated in so many regulatory events is suggested by experimental results found in separate collaborations with the groups of David Clark and Richard Sinden. In the first collaboration, the minimal conditions for *in vitro* transcriptional initiation from the yeast *CUP1* promoter were shown to be a negatively superhelical DNA substrate plus RNAPII; no other regulatory molecules were required to induce transcription from this promoter (Leblanc *et al.*, 2000). Similarly, for replication, the presence of a SIDD site caused by a sufficiently long multimer of the spinocerebellar ataxia type 10 repeat was shown in the second collaboration to cause an apparently uncontrolled replication initiation event (Potman *et al.*, 2003). Together, these results show that the presence of a stress-destabilized region *alone* is sufficient to initiate either replication or transcription. This suggests that stress-induced DNA duplex destabilization could have been a primordial regulator of both of these processes, active at the earliest times in the evolution of DNA regulatory mechanisms. In this scenario, transcription factors, architectural molecules, and other systems would have evolved later to provide more refined levels of control and to specialize regulatory processes to serve a plethora of distinct purposes. Insofar as this evolutionarily ancient role of SIDD in regulation is correct, it suggests that destabilization could have become integrated into many mechanisms of transcriptional or replicational regulation, be they inhibitory or activating.

ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of Jacques Jospitre and Sally Madden in encoding early versions of this algorithm. This work was supported in part by grants DBI 99-04549 from the National Science Foundation and RO1-HG01973 from the National Institutes of Health and by additional support from the Diversa Corporation.

REFERENCES

Amirikyan, B.R., Vologodskii, A.V., and Lyubchenko, Y.L. 1982. Determination of DNA cooperativity factor. *Nucl. Acids Res.* 9, 5469–5482.

Aranda, A., Perez-Ortin, J., Benham, C.J., and del Olmo, M. 1997. Analysis of the *in vivo* structure of a natural alternating d(AT)$_n$ sequence in yeast. *Yeast* 13, 313–326.

Bauer, W.R., and Benham, C.J. 1993. The free energy, enthalpy and entropy of native and of partially denatured closed circular DNA. *J. Mol. Biol.* 234, 1184–1196.

Bauer, W.R., Ohtsubo, H., Ohtsubo, E., and Benham, C.J. 1995. Energetics of coupled twist and writhe changes in closed circular pSM1 DNA. *J. Mol. Biol.* 253, 438–452.

Bauer, W.R., and Vinograd, J. 1970. Interaction of closed circular DNA with intercalative dyes. II. The free energy of superhelix formation in SV40 DNA. *J. Mol. Biol.* 47, 419–435.

Benham, C.J. 1979. Torsional stress and local denaturation in supercoiled DNA. *Proc. Natl. Acad. Sci. USA* 76, 3870–3874.

Benham, C.J. 1980. The equilibrium statistical mechanics of the helix-coil transition in torsionally stressed DNA. *J. Chem. Phys.* 72, 3633–3639.

Benham, C.J. 1981. A theoretical analysis of competing conformational transitions in torsionally stressed DNA. *J. Mol. Biol.* 150, 43–68.

Benham, C.J. 1990. Theoretical analysis of heteropolymeric transitions in superhelical DNA molecules of specified sequence. *J. Chem. Phys.* 92, 6294–6305.

Benham, C.J. 1992. The energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.* 225, 835–847.

Benham, C.J. 1993. Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory regions. *Proc. Natl. Acad. Sci. USA* 90, 2999–3003.

Benham, C.J. 1996a. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.* 255, 425–434.

Benham, C.J. 1996b. Theoretical analysis of the helix-coil transition in positively supercoiled DNA at high temperatures. *Phys. Rev. E* 53, 2984–2987.

Benham, C.J., Kohwi-Shigematsu, T., and Bode, J. 1997. Stress-induced duplex destabilization in chromosomal scaffold/matrix attachment regions. *J. Mol. Biol.* 274, 181–196.

Benham, C.J., Savitt, A., and Bauer, W.R. 2002. Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: Complete determination of energetics based upon a statistical mechanical model. *J. Mol. Biol.* 316, 563–580.

Bloomfield, V., Crothers, D., and Tinoco, I. 1974. *The Physical Chemistry of Nucleic Acids*, Harper and Row, New York.

Bode, J., Kohwi, Y., Dickenson, L., Joh, T., Klehr, D., Mielke, C., and Kohwi-Shigematsu, T. 1992. Biological significance of unwinding capability of nuclear matrix-associating DNA. *Science* 255, 195–197.

Breslauer, K., Frank, R., Bloecker, H., and Marky, L. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83, 3746–3750.

Cheung, K.J., Badarinarayana, V., Selinger, D.W., Janse, D., and Church, G.M. 2003. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic shock response of *Eschicheria coli*. *Genome Res.* 13, 206–215.

Crothers, D.M., and Spatz, H.C. 1971. Theory of friction-limited DNA unwinding. *Biopolymers* 10, 1949–1972.

Dean, W.W., and Lebowitz, J. 1971. Partial alteration of secondary structure in native superhelical DNA. *Nature New Biology* 231, 5–8.

Delacourt, S.G., and Blake, R.D. 1991. Stacking energies in DNA. *J. Biol. Chem.* 266, 15160–15169.

Depew, R., and Wang, J.C. 1975. Conformational fluctuations of DNA helix. *Proc. Natl. Acad. Sci. USA* 72, 4275–4279.

Engels, J.D., and von Hippel, P.H. 1978. Effects of methylation on the stability of nucleic acid conformations. *J. Biol. Chem.* 927–934.

Fye, R.M., and Benham, C.J. 1999. Exact method for numerically analyzing a model of local denaturation in superhelically stressed DNA. *Phys. Rev. E* 59, 3408–3426.

Gelfand, C.A., Plum, G.E., Grollman, A.P., Johnson, F., and Breslauer, K.J. 1998. The impact of an exocyclic cytosine adduct on DNA duplex properties: Significant thermodynamic consequences despite modest lesion-induced structural alterations. *Biochemistry* 37, 12507–12512.

Hatfield, G.W., and Benham, C.J. 2002. DNA topology-mediated control of global gene expression in *Escherichia coli*. *Ann. Rev. Genet.* 36, 175–203.

He, L., Liu, J., Collins, I., Sanford, S., O'Connell, B., Benham, C.J., and Levens, D. 2000. Loss of FBP function arrests cellular proliferation and extinguishes *c-myc* expression. *EMBO J.* 19, 1034–1044.

Huang, R.Y., and Kowalski, D. 1993. A DNA unwinding element and an ARS consensus comprise a replication origin within a yeast chromosome. *EMBO J.* 12, 4521–4531.

Kohwi-Shigematsu, T., and Kohwi, Y. 1990. Torsional stress stabilizes extended base unpairing in suppressor sites flanking immunoglobulin heavy chain enhancer. *Biochemistry* 29, 9551–9560.

Kohwi-Shigematsu, T., and Kohwi, Y. 1991. Detection of triple-helix related structures adopted by poly(dG)-poly(dC) sequences in supercoiled plasmid DNA. *Nucl. Acids Res.* 19, 4267–4271.

Kowalski, D., and Eddy, M.J. 1989. The DNA unwinding element: A novel, *cis*-acting component that facilitates opening of the *E. coli* replication origin. *EMBO J.* 8, 4335–4344.

Kowalski, D., Natale, D., and Eddy, M. 1988. Stable DNA unwinding, not breathing, accounts for single-strand specific nuclease hypersensitivity of specific A+T-rich regions. *Proc. Natl. Acad. Sci. USA* 85, 9464–9468.

Leblanc, B., Benham, C.J., and Clark, D. 2000. An initiation element in the yeast *CUP1* promoter is recognized by RNA polymerase II in the absence of TATA box-binding protein if the DNA is negatively supercoiled. *Proc. Natl. Acad. Sci. USA* 97, 10745–10750.

Lerman, L.S., and Silverstein, K. 1987. Computer simulation of DNA melting and its application to denaturing gradient gel electrophoresis. *Methods Enzymol.* 155, 482–501.

Lilley, D.M.J. 1980. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc. Natl. Acad. Sci. USA* 77, 6468–6472.

Liu, L.F., and Wang, J.C. 1987. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA* 84, 7024–7027.

Murchie, A.I.H., and Lilley, D.M.J. 1989. Base methylation and local DNA helix stability. *J. Mol. Biol.* 205, 593–602.

Parvin, J.D., Shykind, B.M., Meyers, R.E., Kim, J., and Sharp, P.A. 1994. Multiple sets of basal factors initiate transcription by RNA polymerase II. *J. Biol. Chem.* 269, 18414–18421.

Peck, L.J., Nordheim, A., Rich, A., and Wang, J.C. 1982. Flipping of cloned $d(pCpG)_n \cdot d(pCpG)_n$ DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. *Proc. Natl. Acad. Sci. USA* 79, 4560–4564.

Potaman, V., Bissler, J., Hashem, V., Oussatcheva, E., Lu, L., Shlyakhtenko, L., Lybuchenko, Y., Matsuura, T., Ashizawa, T., Leffak, M., Benham, C.J., and Sinden, R.R. 2003. Unwound structures in SCA10 $(ATTCT)_n \cdot (ATTCT)_n$ repeats. *J. Mol. Biol.* 326, 1095–1111.

Pulleyblank, D., Shure, M., Tang, D., Vinograd, J., and Vosberg, H.-P. 1975. Action of nicking-closing enzyme on supercoiled and non-supercoiled closed circular DNA: Formation of a Boltzmann distribution of topoisomers. *Proc. Natl. Acad. Sci. USA* 72, 4280–4284.

Rahmouni, A.R., and Wells, R.D. 1989. Stabilization of Z DNA *in vivo* by localized supercoiling. *Science* 246, 358–363.

Rothman-Denes, L.B., Dai, X., Davydova, E., Carter, R., and Kazmierczak, K. 1998. Transcriptional regulation by DNA structural transitions and single-stranded DNA-binding proteins. *Cold Spring Harbor Symp. Quant. Biol.* 63, 63–73.

Schildkraut, C., and Lifson, S. 1968. Dependence of the melting temperature of DNA on the salt concentration. *Biopolymers* 3, 195–208.

Sheridan, S., Benham, C.J., and Hatfield, G.W. 1998. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoil-induced DNA duplex destabilization in an upstream activating sequence. *J. Biol. Chem.* 273, 21298–21308.

Sheridan, S., Benham, C.J., and Hatfield, G.W. 1999. Inhibition of DNA supercoiling-dependent transcriptional activation by a distant B-DNA to Z-DNA transition. *J. Biol. Chem.* 274, 8169–8174.

Singleton, C.K., Klysik, J., Stirdivant, S.M., and Wells, R.D. 1982. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature* 299, 312–316.

Steger, G. 1994. Thermal denaturation of double-stranded nucleic acids. *Nucl. Acids Res.* 22, 2760–2768.

Sun, H.-Z., Mezei, M., Fye, R.M., and Benham, C.J. 1995. Monte Carlo analysis of conformational transitions in superhelical DNA. *J. Chem. Phys.* 103, 8653–8665.

Tal, M., Shimron, F., and Yagil, G. 1994. Unwound regions in yeast centromere IV DNA. *J. Mol. Biol.* 243, 179–189.

Vesnaver, G., Chang, C.-N., Eisenberg, M., Grollman, A., and Breslauer, K. 1989. Influence of abasic and anucleosidic sites on the stability, conformation and melting behavior of a DNA duplex: Correlations of thermodynamic and structural data. *Proc. Natl. Acad. Sci. USA* 86, 3614–3618.

Vinograd, J., Lebowitz, J., and Watson, R. 1968. Early and late helix-coil transitions in closed circular DNA. *J. Mol. Biol.* 33, 173–197.

Wang, J.C. 1979. Helical repeat of DNA in solution. *Proc. Natl. Acad. Sci. USA* 76, 200–203.

Woodworth-Gutai, M., and Lebowitz, J. 1976. Introduction of interrupted secondary structure in supercoiled DNA as a function of superhelix density: Consideration of hairpin structures in superhelical DNA. *J. Virol.* 18, 195–204.

Woynarowski, J.M., Trevino, A., Rodriguez, K., Hardies, S.C., and Benham, C.J. 2001. AT-rich islands in genomic DNA—A novel target for AT-specific DNA-reactive antitumor drugs. *J. Biol. Chem.* 276, 40555–40566.

Yamaki, H., Ohtsubo, E., Nagai, K., and Maeda, Y. 1988. The *oriC* unwinding by *dam* methylation in *Escherichia coli*. *Nucl. Acids Res.* 16, 5067–5073.

Address correspondence to:
Craig J. Benham
UC Davis Genome Center
University of California
One Shields Avenue
Davis CA 95616

E-mail: cjbenham@ucdavis.edu