*J. Mol. Biol.* (1996) **255**, 425–434

JMB

# Duplex Destabilization in Superhelical DNA is Predicted to Occur at Specific Transcriptional Regulatory Regions

**Craig J. Benham**

*Department of Biomathematical Sciences Box 1023, Mount Sinai School of Medicine, 1 Gustave Levy Place, New York, NY 10029, USA*

Analytic methods that accurately calculate the extent of duplex destabilization induced in each base-pair of a DNA molecule by superhelical stresses are used to analyze several genomic DNA sequences. Sites predicted to be susceptible to stress-induced duplex destabilization (SIDD) are found to be closely associated with specific transcriptional regulatory regions. Operators within the promoters of SOS-regulated *Escherichia coli* genes are destabilized by superhelical stresses, whereas closely related sequences present elsewhere on that genome are not. Analysis of genomic sequences from the budding yeast *Saccharomyces cerevisiae* finds a distinctive tripartite pattern, in which the 3' and 5' termini of genes are destabilized, but the sequence encoding the primary transcript is not. Three rDNA genes from higher eukaryotes exhibit a similar pattern. Implications of these results regarding possible mechanisms of activity of the regions involved are discussed. A strategy is presented for designing experiments in which the susceptibility to SIDD of a local region is altered without changing its local base sequence. The occurrence of the observed SIDD patterns provides a new approach to searching uncharacterized genomic sequences for transcriptionally active regions.



*Keywords:* DNA denaturation; transcriptional regulation; SOS regulation; DNA superhelicity; duplex stability

## Introduction

Local DNA denaturation is an obligatory step in the initiation of transcription and of replication, and may also play roles in other DNA functions. For this reason one expects the locations and occasions of its occurrence to be stringently controlled. DNA superhelicity, which is closely regulated *in vivo*, can induce the formation of locally unpaired regions at defined sites within DNA molecules (Dean & Lebowitz, 1971; Beerman & Lebowitz, 1973). Nuclease digestion experiments have shown this local denaturation to occur at specific regulatory regions. In pBR322 DNA it is confined to two locations, the 3' terminus of the β-lactamase gene, and the promoter region of the same gene (Kowalski *et al*., 1988). The *E. coli oriC* replication origin also undergoes superhelical melting at a precise location (Kowalski & Eddy, 1989). When the base sequence of this site is altered, replication occurs *in vivo* only if susceptibility to stress-induced denaturation at the correct position is retained. Other attributes, including sequence homology, do not affect activity. Although mechanisms of DNA function commonly involve complex interactions with other molecules, the intrinsic susceptibility to stress-induced base-unpairing at specific sites clearly is essential for activity.

This attribute is not determined by strictly local properties of DNA sequence, but rather by a global competition among all sites within a stressed domain, be it a circular molecule or a linear segment bounded by anchors. The opening of any single base-pair alters its helical twist, changing the distribution of the superhelical deformation (i.e. the linking difference) and thereby affecting the denaturation probability of every other base-pair in the domain. In this way superhelicity creates a global coupling among the conformational states of all base-pairs.

Theoretical methods have been developed to analyze denaturation in superhelical DNA domains of specified sequence (Benham, 1990). A statistical mechanical approach is used, in which the governing partition function and other quantities of

Abbreviations used: SIDD, stress-induced duplex destabilization; ORFs, open reading frames.

0022–2836/96/030425–10 $12.00/0

interest are evaluated approximately to a high degree of precision. The free-energy parameters governing this transition have been evaluated from the results of endonuclease digestion experiments on superhelical pBR322 DNA (Kowalski *et al*., 1988; Benham, 1992). When these parameter values were used in analyses of superhelical denaturation in other molecules for which experimental data were available, in all cases the predicted locations and extents of denaturation agreed precisely with those observed, within the limits of experimental accuracy (Benham, 1992). Because these theoretical methods have been shown to depict the superhelical denaturation observed to occur in diverse DNA domains in a quantitatively accurate manner, they may be used to predict this potentially important physical chemical attribute in other DNA sequences.

Imposed stresses can influence biological activity by more subtle processes than the driving of local denaturation. Sub-threshold destabilization of the DNA duplex may also be important, that is, imposed stresses may decrease the free energy required to open a local site, but not by enough to induce its denaturation. This phenomenon will be important when duplex opening occurs by processes, perhaps involving interactions with other molecules, that can provide sufficient free energy to cause local melting only if the DNA site involved already is marginally destabilized by stresses. An extension of the original theoretical technique permits the evaluation of stress-induced duplex destabilization (SIDD) experienced throughout a DNA domain due to imposed superhelicity (Benham, 1993).

Preliminary analyses have shown that the sites predicted to experience SIDD do not occur at random, but instead are found at specific regulatory regions (Benham, 1993). Here we analyze the destabilization behavior of numerous superhelically stressed genomic DNA sequences containing transcriptionally active regions.

## The analysis of stress-induced destabilization

Because less free energy is needed to induce opening of A·T base-pairs than of G·C pairs at physiological temperatures and ionic strengths (Breslauer *et al*., 1986; Delacourt & Blake, 1991), predictions of the locations of stress-induced denaturation in principle could be based on either local A + T-richness or local unpairing energy (Natale *et al*., 1992). While this approach may correctly indicate which sites are most susceptible to opening, evaluating the extent of superhelical destabilization throughout a domain requires a more complete analysis. Whether a site melts under defined superhelical conditions depends not just on its local attributes, but also on its interactions with all other sites in the domain. This coupling induces complex transition behavior, as illustrated by the results presented below.

A complete description of the statistical mechanical method for analyzing superhelical duplex destabilization has been presented elsewhere (Benham, 1990, 1992). Briefly, a free energy is associated with each state of base-pairing of the DNA that depends on three factors—the number and base compositions of the unpaired regions, the extent of interstrand twisting experienced by these regions, and the superhelical deformation (i.e. linking difference) of the DNA domain involved. The first step in the analysis is to determine the state of minimum free energy. Then a threshold energy value is specified, and all states are found whose free energy exceeds that of the minimum energy state by no more than this threshold amount. Initial expressions for the partition function and other important statistical mechanical quantities are evaluated from these states. The contributions from all the states which do not satisfy the energy threshold condition are estimated by a density of states procedure. Its results are used to refine the initial expressions determined above to account for the influence of these high-energy states. This last step, although approximate, has been shown to be highly accurate. Deviations of any refined quantity from its exact value do not exceed 0.01% (Benham, 1990).

This analysis calculates two quantities which describe the stability properties of the sequence. First, the ensemble average probability $p(x)$ of melting of the base-pair at position $x$ in the sequence is given by:

$$p(x) = \frac{Z(x)}{Z}$$

where $Z$ is the partition function and:

$$Z(x) = \sum_{i_x} \exp[-G(i_x)/RT]$$

This summation is performed over all states $i_x$ in which the base-pair at position $x$ is open. (For simplicity, all states will be denoted here as though they are discrete in character. For parameters that vary continuously it is understood that relevant summations actually involve integrals.) The graph of $p(x)$ *versus* $x$ is called the transition profile, and delineates those regions of the superhelical domain where duplex opening occurs with a significant probability.

A more sensitive measure of destabilization is found by calculating the incremental energy $G(x)$ needed to separate the base-pair at position $x$ (Benham, 1993). This quantity is calculated as:

$$G(x) = \bar{G}(x) - \bar{G}$$

where $\bar{G}$ is the ensemble average free energy of the system and $\bar{G}(x)$ is the average free energy of all states $i_x$ in which the base-pair at position $x$ is separated:

$$\bar{G}(x) = \frac{\Sigma_{i_x} G(i_x) \exp[-G(i_x)/RT]}{Z(x)}$$

Here $G(i_x)$ is the free energy of the state $i_x$. Stress-induced duplex destabilization (SIDD) profiles are plots of $G(x)$ *versus* $x$. SIDD profiles are more informative than transition profiles because they also depict sites of sub-threshold destabilization, where the amount of free energy needed to induce duplex opening is decreased relative to neighboring regions.

## The global interactive character of superhelical destabilization

The global coupling induced by superhelicity can produce complex transition behavior. The probability of denaturation of each base-pair need not increase monotonically as the level of destabilizing stress rises. Instead, as superhelical stresses increase, regions that initially were denatured may re-anneal as other sites open elsewhere in the molecule. A coupled transition of this type is illustrated in Figure 1, which shows denaturation probability profiles of the chicken histone H5 gene region at two linking differences. A virtually complete reversal of transition occurs between the two most sensitive sites as the imposed stress is increased. This behavior occurs when the relative competitiveness of sites susceptible to denaturation varies with the amount of stress imposed. No static descriptor of local helix stability, such as the A + T-richness profile or local opening energy, can accurately predict this non-monotonic, interactive behavior.

The importance of molecular context in determining SIDD behavior is illustrated in Figure 2. Part (a) depicts the SIDD profile of the yeast CYC1 gene region. The sequence encoding the primary transcript is marked with a bar, and transcribes towards the right. The only sites experiencing significant SIDD occur at the beginning and the end of the transcribed region, with the 3' terminus being most destabilized. (This pattern is seen in other yeast genes, as described below.) Deletion of the 38 base-pairs between the vertical bars in the 3'-terminal region yields the sequence whose SIDD profile under identical conditions is shown in Figure 2(b). This deletion removes part of the site that dominates destabilization in the intact molecule, thereby substantially decreasing its competitiveness. Because the mutated sequence in this case contains no other dominant site, a large number of approximately equally competitive sites become destabilized. The massive changes of duplex stability seen throughout the sequence show the strong effect on global transition behavior of the presence or absence of small numbers of base-pairs at strategic locations. This also cannot be predicted from local sequence attributes, which are altered only at the deletion site.

## SIDD at transcriptional regulatory regions

In this section we report the results of calculations of the SIDD profiles of several genes, together with their transcriptional regulatory regions. The sequences analyzed were either complete entries in the GenBank database (Bilofsky, *et al.*, 1986), or were extracts from a larger context, such as the sequence of a complete chromosome. In the latter case the sequence extract that was analyzed was chosen to comprise the entire site of interest, plus flanking regions. This approach allowed the destabilization properties of these transcriptional units to be analyzed in a uniform manner. Sample calculations demonstrate that the destabilization properties of the region of interest do not vary greatly with the size of the extracted sequence being analyzed; (data not shown.) Only when a larger extracted sequence contains a highly competitive extraneous site that is excluded from the smaller extracted sequence is there a

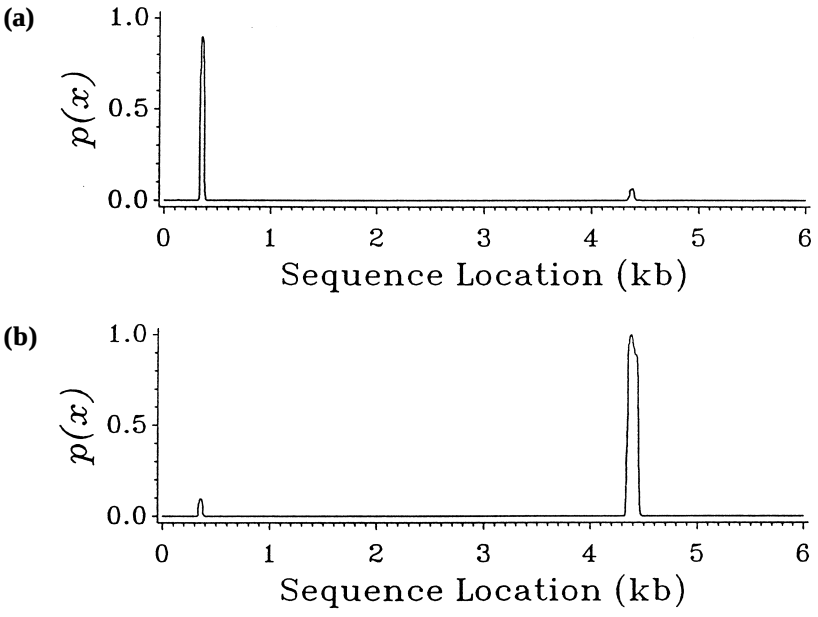


**Figure 1.** The probability of denaturation is plotted as a function of base-pair position for the genomic region containing the chicken histone H5 gene sequence at two values of linking difference, (a) $\Delta Lk = -27$ turns, and (b) $\Delta Lk = -35$ turns. These correspond to superhelix densities of $\sigma = -0.047$ and $\sigma = -0.060$, respectively. The site near position 350 whose denaturation dominates at $\Delta Lk = -27$ has almost completely reverted to the duplex form when $\Delta Lk = -35$. This reversion is coupled to transition near base-pair 4400.

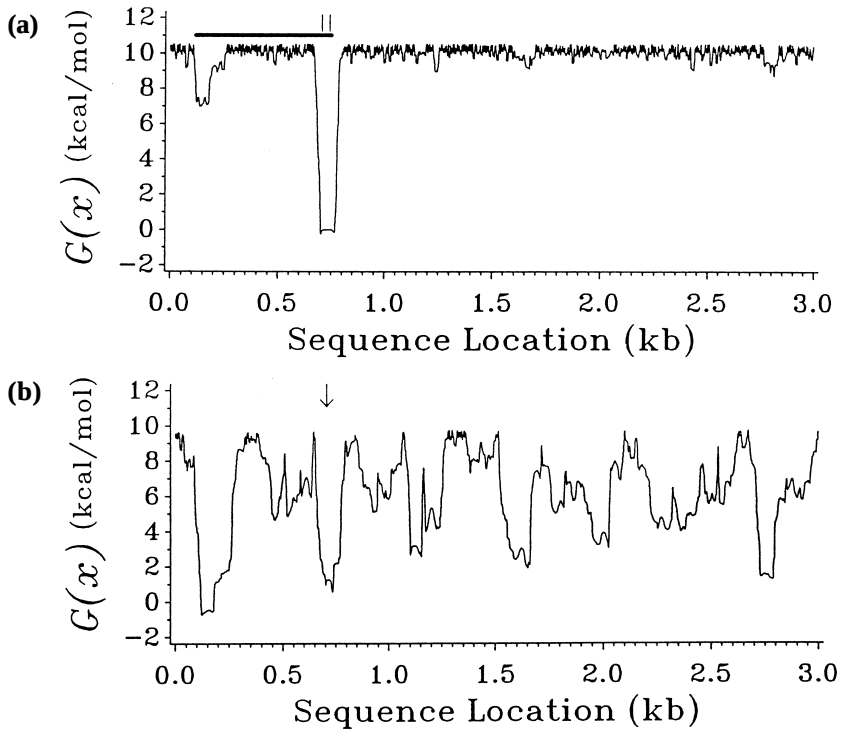


**Figure 2.** The destabilization profile at $\Delta Lk = -18$ turns of a 3 kb yeast genomic sequence containing the CYC1 gene is shown in (a). This corresponds to a superhelix density of $\sigma = -0.063$. The primary transcript coding region is denoted by the horizontal line. When the 38 base-pairs located between the vertical bars within the CYC1 3' terminal region are deleted, the transition profile changes to that shown in (b), where the arrow marks the deletion site.

quantitative change in the extent of destabilization of the region of interest. However, even in these cases the relative amounts of destabilization experienced by different parts of that region remain almost invariant. In no case is the small amount of variation observed sufficient to alter any of the conclusions of our analysis.

All calculations reported here use the energy parameter values appropriate for detecting superhelical denaturation by the nuclease digestion procedure of Kowalski *et al*. (1988; Benham, 1992).

## SOS-regulated genes in *E. coli*

The ColE1 plasmid DNA sequence contains a SIDD site at the *cea-kil* operon promoter region (Benham, 1993). Expression of this SOS-regulated operon is repressed by the binding of LexA to the operator region of its promoter. The site of maximum predicted destabilization within the promoter region coincides precisely with the location of this operator. This concurrence suggests that SIDD might be an important attribute of SOS-regulated promoters.

To test this hypothesis, destabilization profiles were calculated for all the genes of the SOS system whose sequences, including their upstream flanker regions, were found in the GenBank database. Suitable sequence information was available for eight genes or operons known to be regulated by LexA (Little & Mount, 1982; Benson *et al*., 1988). These are the six genomic sequences *umuCD, uvrA, lexA, dinA, recA* and *ruvAB*, and the plasmid operon sequences *mucAB* on the pKM101 plasmid and *cea-kil* on ColE1. Figure 3 shows the SIDD profile for the *uvrA* gene sequence.

In order to investigate SIDD under equivalent conditions in sequences of different lengths, the standard was adopted to examine the first linking difference at which the probability of the entirely *B*-form state fell below $1 \times 10^{-4}$. The base-pair at position $x$ was regarded as destabilized if the

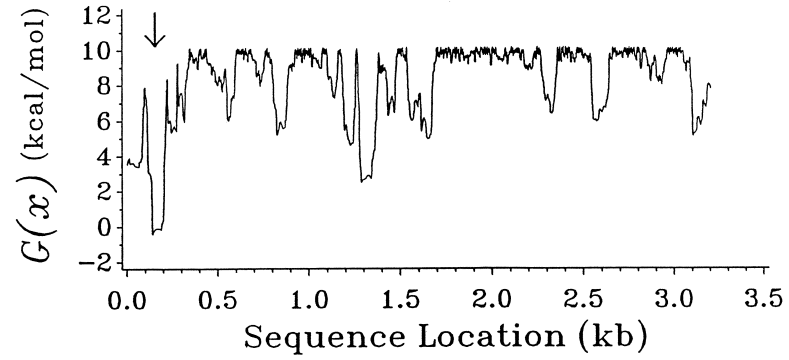


**Figure 3.** The SIDD profile of the genomic region containing the *uvrA* gene is shown. The arrow shows the location of the LexA binding site.

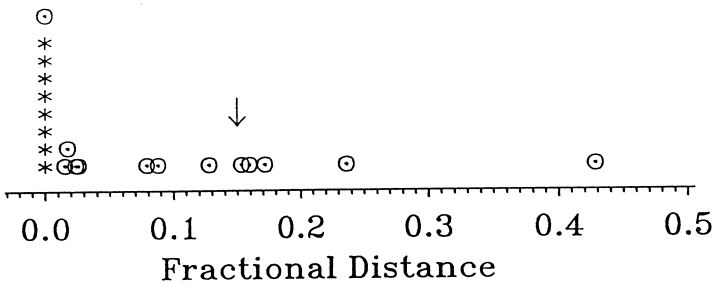


**Figure 4.** The distances from each LexA binding site to the nearest destabilized site are shown by asterisks for the eight SOS-regulated operons described in the text. Circles denote the analogous distances for the 13 sequences containing sites highly homologous to the *lexA* operator. Distances have been normalized as fractions of the sequence length. The average fractional distance that would be expected if no correlation existed between consensus sequence locations and SIDD sites is marked with an arrow.

incremental free energy $G(x)$ needed to induce opening there fell below 5 kcal/mol. Each set of contiguous destabilized base-pairs constituted a destabilized region. All the destabilized regions in a given sequence were found, and ranked according to the minimum value attained by $G(x)$ within each. Next, the LexA binding sites were found, either from annotations in the database or by homology with the consensus binding sequence CTG-TATAW(5)CAG. (Here W denotes either A or T, and (5) stands for five base-pairs of unrestricted type.) The distance between this binding site and each of the destabilized regions was evaluated as the number of base-pairs intervening between them. In all cases the LexA binding site was found to occur in the interior of one of the most destabilized regions in the sequence. In four cases this was the most destabilized region, in three it was the second most destabilized region, and in one case it was the third most destabilized region. Moreover, in every case the LexA binding site occurred at the most destabilized portion of the region involved.

Next, we tested whether SIDD was associated simply with the base sequence of the site involved, or whether it was specifically restricted to those sequences which occurred within promoter regions. We extracted from the database all *E. coli* sequences containing a site with sequence CTGTWTNW(4,6)-CAG, which slightly generalizes the consensus LexA binding sequence. (Here (4,6) denotes from four to six base-pairs of unrestricted type.) After excluding entries in which this local sequence occurred at a promoter, entries shorter than 500 base-pairs (the lower limit for an informative analysis), and the known SOS-regulated operons, 13 sequences remained. Each contained between three and five SIDD sites, determined as described above. The distance from the consensus sequence to the nearest SIDD site was found in each case. Plots of these distances are given in Figure 4, both for the eight SOS-regulated operons and for the population of 13 similar sequences generated here. Because the sequences varied in length, these results were standardized by computing separation distances as fractions of sequence length. Only one of these 13 sequences was destabilized at the consensus site. This occurred at position 1136 to 1151 of the ECOHU43 sequence, in the terminal flanking region of the *yeeC* gene. However, 3' flanking regions are among the most destabilized sites present in genomic sequences (see below), so destabilization of this site may be attributable to its terminal flanking character rather than to the presence of the consensus sequence.

To evaluate whether the distances found by this procedure are randomly distributed, we determined the average fractional separation distance expected in the random case. In the unit interval [0, 1] we chose a random point $p$ and three other random points $q_1$, $q_2$ and $q_3$, then found the distance from $p$ to the nearest $q_i$. When this was done $1 \times 10^6$ times, the average distance was found to be 0.1498 (denoted by an arrow in Figure 4). The average fractional distance between the consensus site and the nearest destabilized region in our sample of 13 sequences was 0.117. The difference between these values is not statistically significant, so we conclude that there is no association between destabilized sites and those locations having strong sequence similarity to the LexA binding site that do not occur in promoter regions. Thus the observed destabilization of SOS-regulated operator regions is not attributable to the sequence of the LexA binding site *per se*. Rather, it correlates exactly with the property that the region involved is an SOS-regulated promoter.

The search for sites whose sequences are similar to the LexA binding sequence discovered two locations that coincide with promoters. These occur at positions 7091 to 7106 and 42,695 to 42,710 in the ECOUW82 sequence of the region between 81.5 and 84.5 minutes on the *E. coli* genome. Both sequences are approximate palindromes, and are highly similar to known LexA binding sites when read in either direction. They both are exactly 16 base-pairs long, as are LexA binding sites, and have the correct bases at all highly conserved positions. The SIDD profiles computed for regions containing 5000 base-pairs centered on each of these sites show that both locations coincide precisely with highly destabilized regions (data not shown), as do the operators of all SOS-regulated genes examined. This suggests that one or both of these promoters could be regulated by the SOS system. At least one

LexA-repressed gene is present in the 80 to 85 minute region of the *E. coli* genome, although its precise location is not known (Kenyon & Walker, 1980; Little & Mount, 1982).

The observation that SIDD at the LexA binding site is an attribute of all SOS-regulated promoters examined suggests it may be involved in their mechanisms of activity. One possibility is suggested by studies showing that the onset of anaerobicity induces an increase in chromosomal DNA superhelicity (Yamamoto & Droffner, 1985), which activates *cea* gene transcription (Malkhosyan *et al*., 1991). Decreasing the affinity of LexA for its binding site, as would occur were it unable to bind to a denatured operator, will increase the frequency of transcriptional initiation. Moreover, if RecA, which binds readily to single-stranded DNA, were to occupy the SIDD site, it could maintain transcriptional activity until it was released and the operator returned to the duplex state. This proposed mechanism requires SIDD at the operator of any SOS-inducible promoter that is activated in response to anaerobicity. If SIDD is required for function, then decreasing or eliminating the susceptibility of the LexA operator region to denature when stressed would inhibit or eliminate SOS activation of the promoter involved. A strategy for designing experiments of this type is described in the Discussion section below.

## Destabilization patterns around eukaryotic genes

Initial calculations on eukaryotic viral DNA sequences found 3' termini of coding regions to be the most susceptible sites to SIDD on the molecules analyzed (Benham, 1993). To examine this association further, we have analyzed the SIDD profiles of several transcribed regions from the yeast *Saccharomyces cerevisiae*.

We initially determined the SIDD profiles of ten yeast genomic DNA sequences containing the following 13 genes: ADH1, ADH2, CYC1, FBP1, GAL7, PHO3, PHO5, POT1, SUC2, TRP1, TRP3, URA3 and UTR1. These sequences were selected by J. Perez-Ortin as being representative of the active regions of the yeast genome. A second set of gene sequences also was analyzed, consisting of the following 13 genes from chromosome III: ABP1, BIK1, CHA1, CIT2, CRY1, DTP1, GLK1, HIS4, PDI1, PET18A, PET18B, PGK1 and THR4. Figure 5 plots the calculated SIDD profiles for five representative sequences containing seven genes. In analyzing these results we regard a site as significantly destabilized if its calculated denaturation energy $G(x)$ is at least 2 kcal/mol less than that of the baseline.

The calculated SIDD profiles show a clear pattern around the coding regions consisting of three components. The strongest component, present in every sequence examined, is a site of strong destabilization in the 3' terminal flanking region. Second, the sequence encoding the primary transcript contains no destabilized regions. This occurs in 21 of the 26 genes examined. The third component of the pattern, destabilization in the 5' flanking region of the gene, is found in 20 genes. Commonly 5' flanks are less destabilized than are their corresponding 3' terminal regions. Figure 5(e) shows the SIDD profile for the genomic region encoding the genes HIS4 and BIK1, which are representative of the observed exceptions to the pattern. HIS4 contains an internal region near its 5' end that is destabilized by approximately 3 kcal/mol, violating the second part of the pattern. Although the 5' region of the BIK1 gene is slightly destabilized, this is not enough for it to satisfy the third component of the pattern.

SIDD profiles also have been calculated for the RNA *Pol*I-transcribed rDNA sequences from three organisms, *Tetrahymena*, *Drosophila*, and mouse. Figure 6 shows the *Tetrahymena* SIDD profile. In all three cases the primary transcript is bracketed by destabilized sites at or near its 5' and 3' ends. The interior of the transcribed region contains a destabilized site only in *Drosophila*.

One possible explanation for the tripartite pattern found in all the sequences treated here involves the effect of the polymerase transcription complex on its DNA substrate. Translocation of this complex pushes a wave of positive supercoils ahead, and leaves a wake of negative supercoils behind (Liu & Wang, 1987; Tsao *et al*., 1989). Progress may be impeded by an accumulation of positive supercoils in the downstream regions. A denatured site near the 3' transcription terminus could act as a sink to absorb the positive supercoils generated by translocation, thereby increasing the efficiency of transcription from the gene involved. This also would isolate that gene from its downstream neighbor by preventing any positive supercoils generated by its transcription from propagating beyond its 3' flank. The wake of negative supercoils could destabilize the promoter region of the gene, thereby facilitating the next round of transcription. The presence of SIDD sites internal to a gene would impede both these processes. This hypothetical scenario can account for all three attributes of the detected pattern. It does not require that all transcribed regions experience the same amount of downstream destabilization, nor that the SIDD sites be precisely positioned relative to the 3' termini of the primary transcripts. The overall rate of translocation of the transcription complex may be controlled by other factors, with the destabilized downstream site required only to remove a geometric impediment to progress.

SIDD around the 3' terminal region of a gene may be involved in the mechanisms by which the primary transcript is terminated. Support for this explanation comes from an experiment in which 38 base-pairs in the 3' terminal region of the yeast CYC1 gene were deleted (Zaret & Sherman, 1982). This deletion causes massive changes in the predicted SIDD profile, shown in Figure 2 above, greatly decreasing stability throughout the entire

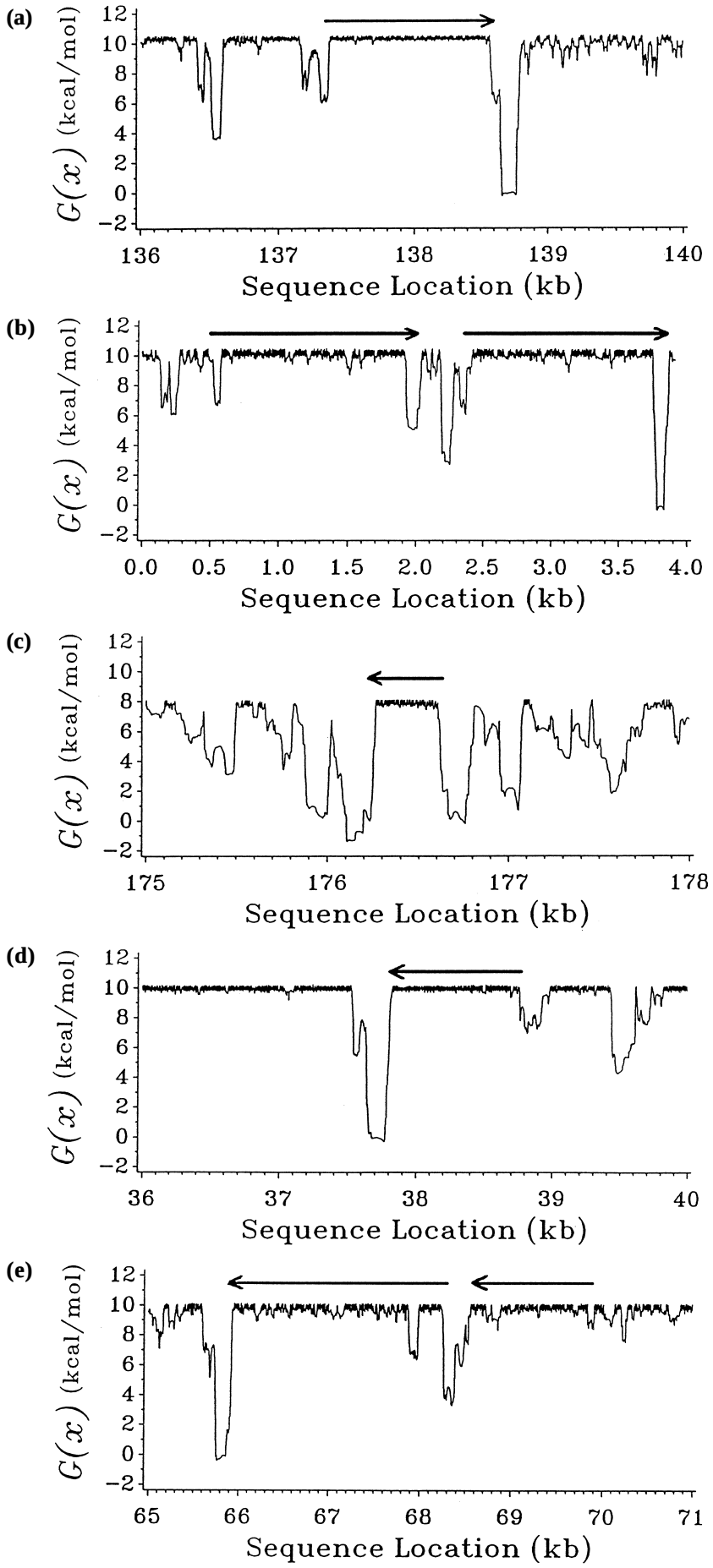


**Figure 5.** SIDD profiles are shown for five representative yeast genomic sequences of yeast; (a) the PGK1 gene, (b) the PHO5 (left) and PHO3 (right) genes, (c) the CRY1 gene, (d) the DTP1 gene, (e) the HIS4 (left) and the BIK1 (right) genes. In each case the coding region is denoted by an arrow.

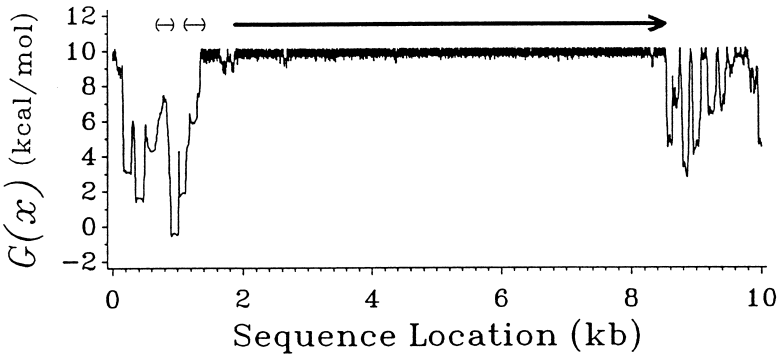


**Figure 6.** The SIDD profile of the *Tetrahymena* rDNA gene sequence is shown. The region corresponding to the primary transcript is denoted by an arrow. This sequence occurs *in vivo* as a palindromic dimer. The profile shown is for one copy of this dimer, with the axis of symmetry occurring at position zero.

sequence. Figure 7 shows the transition probability profiles calculated for the wild-type (continuous line) and the mutated (broken line) sequences. This deletion decreases the probability of denaturation of the 3' terminus of the CYC1 gene from unity to 0.12. Experiments have shown it also causes a corresponding change in the site of transcription termination (Zaret & Sherman, 1982). Whereas termination in the wild-type sequence occurs exclusively within the destabilized site around position 730, termination in the mutated sequence occurs at this normal position approximately 10% of the time. The deletion did not remove any sequence essential for termination, as the normal event remained possible. However, the observed decrease in the frequency of normal termination was comparable to the decrease in the calculated frequency of duplex opening of the 3' region, suggesting that local denaturation may play some role in the events by which transcription is terminated.

## Discussion

This work reports the predicted duplex destabilization characteristics of diverse superhelically stressed genomic DNA sequences, as found using recently developed analytic methods of demonstrated accuracy. This newly calculable physical chemical attribute—the susceptibility to SIDD of sites within a DNA domain—is found to be closely associated with specific types of regulatory regions. The strengths of the associations found here suggest that SIDD may play roles in several types of physiological activities. Although our results provide insight into possible mechanisms of function, the precise role of SIDD in each case must be determined experimentally.

The global coupling induced by superhelicity can be exploited to design a class of experiments in which the susceptibility to SIDD is altered at a specific site by changing the base sequence at other locations. Insertion or removal at a remote location of a sequence susceptible to destabilization will change the efficiency with which the opening transition at the modified site competes with opening at the site of interest in a superhelical domain, altering its probability of denaturation accordingly. In principle this allows the design of experiments in which one alters the susceptibility to SIDD at a specific site without changing its local sequence. Experiments of this type could be designed to probe the role of SIDD in specific regulatory events.

We report a tripartite SIDD pattern occurring at yeast genes, in which the region encoding the primary transcript is not destabilized, but has destabilized sites in its 5' and 3' flanks. This pattern requires that flanking regions of genes have higher average A + T content than their coding regions. Recent analysis of the complete sequences of yeast chromosomes III and XI has found that average G + C content is highest for open reading frames (ORFs), somewhat smaller for divergent putative promoters (assuming the ORFs are genes) and smaller still for convergent putative terminators (Dujon *et al*., 1994). The present work suggests that these compositional variations may be explained by a requirement to preserve the tripartite SIDD pattern around transcribed regions. To perpetuate

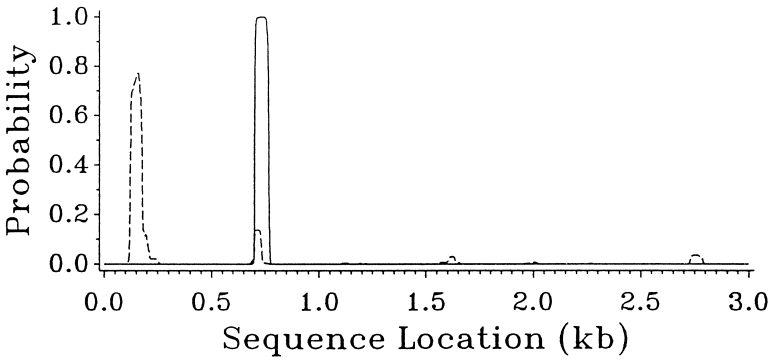


**Figure 7.** Transition profiles for the yeast CYC1 gene region are shown. The profile for the wild-type sequence is denoted with the continuous line, and the profile of the sequence with the 38 bp deletion described in Figure 2 is given by the broken line.

this pattern, codon selections must maintain higher local G + C content in coding regions than in their flanks, producing a selection pressure away from maximal A + T-richness. Yet a tendency has been documented towards using A + T-rich codons in maximally expressed yeast genes (Sharp & Cowe, 1991). If the tripartite pattern found here has no functional significance, then this trend towards A + T-richness over time should have expunged it.

We note that this pattern may not be common to all yeast genes, despite finding it in all sequences examined here. By analyzing only known genes, an unavoidable selection bias is introduced towards genes whose products are either abundant or important. It is possible that other genes, excluded from this group, might not exhibit this pattern.

Fast and accurate computational methods are needed to search experimentally uncharacterized DNA sequences for coding and regulatory regions. The close associations reported here between specific patterns of destabilization and particular types of transcriptionally active regions could in principle be incorporated into algorithms for performing these types of searches. A first example of such an application was given in the analysis of SIDD at LexA repressor binding sites of SOS-regulated genes. Two additional promoters were found that contain SIDD sites at sequences homologous to the *lexA* operator. These occur in a region of the *E. coli* genome known to contain at least one uncharacterized SOS-regulated gene.

The tripartite SIDD pattern found in yeast genes may be incorporated into algorithms which search that genome for coding regions. The strength of this association suggests that this procedure could yield significant increases in the sensitivity and specificity achieved. This question will be addressed in a future paper.

The stresses which induce duplex destabilization are regarded here as arising from superhelicity. This is the standard experimental method for imposing stresses on DNA, and it has clear biological importance. However, there are many other ways that stresses on DNA can be modulated *in vivo*. The binding of proteins or other molecules may affect the state of stress of the DNA substrate. Changes in the structure or position of nucleosomes could alter stresses within DNA linker regions. Disassociation of nucleosomes from DNA can induce stresses by converting restrained supercoils into unrestrained ones. Translocation of the RNA polymerase transcription complex pushes a wave of positive supercoils ahead, and leaves a wake of negative supercoils behind, altering the state of tension of the DNA in these regions (Liu & Wang, 1987; Tsao *et al*., 1989). In principle, any deformation which alters the torsional stresses imposed on DNA in a region can affect duplex stability there in qualitatively similar ways to those described here.

Our understanding of the precise manner in which superhelicity and other factors may stress DNA *in vivo* is complicated by its associations with other molecules, primarily nucleosomes or HU proteins. Because nucleosomal winding restrains DNA tertiary structure, any stresses must be primarily torsional in character. A simple theoretical analysis suggests that conformational transitions may be driven at less extreme linking differences when DNA tertiary structure is restrained (Benham, 1987). How far these stresses propagate in chromatin is not known. If they are damped out over a distance, their influence would be confined to a local region (Umek & Kowalski, 1990). In this case the class of experiments described above, in which the stability properties of regions are modified by sequence alterations at remote sites, may only work *in vivo* if the separation distance between the sites involved is sufficiently small. Such damping also could allow stresses to be imposed on regions within unanchored linear molecules that do not constitute topological domains.

This work documents clear associations between SIDD susceptibility and specific transcriptionally active regions. Many other types of genomic regulatory regions and binding sites currently are being examined to determine their SIDD properties. The results of this work will be reported in future papers.

## Acknowledgements

I gratefully acknowledge the kind assistance of Dr Jose Perez-Ortin in selecting yeast sequences for analysis, and of Joshua Lederberg and David Thaler for their thoughtful suggestions and comments on the manuscript. This research was supported in part by grants RO1-GM-47012 from the National Institutes of Health and BIR 93-10252 from the National Science Foundation. Use of the GenBank Database in the Wisconsin Sequence Analysis Package (GCG) was supported in part by grant 5-MO1-RR00071 to the Mount Sinai General Clinical Research Center from the Center for Research Resources of the National Institutes of Health.

## References

Beerman, T. A. & Lebowitz, J. (1973). Further analysis of the altered secondary structure of superhelical DNA. Sensitivity to methylmercuric hydroxide, a chemical probe for unpaired bases. *J. Mol. Biol.* **79**, 451–470.

Benham, C. J. (1987). The influence of tertiary structural restraints on conformational transitions in superhelical DNA. *Nucl. Acids Res.* **15**, 9985–9995.

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). 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 loci. *Proc. Natl Acad. Sci. USA*, **90**, 2999–3003.

Benson, F. E., Illing, G. T., Sharples, G. J. & Lloyd, R. G. (1988). Nucleotide sequencing of the *ruv* region of

*Escherichia coli* reveals a LexA regulated operon encoding two genes. *Nucl. Acids Res.* **16**, 1541–1549.
Bilofsky, H. S., Burks, C., Fickett, J. W., Goad, W. B., Lewitter, F. I., Rindone, W. P., Swindell, C. D. & Tung, C.-S. (1986). The GenBank genetic sequence databank. *Nucl. Acids Res*., **14**, 1–4.
Breslauer, K. J., Frank, R., Bloecher, H. & Marky, L. A. (1986). Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
Dean, W. W. & Lebowitz, J. (1971). Partial alteration of secondary structure in native superhelical DNA. *Nature New Biol.* **231**, 5–8.
Delacourt, S. G. & Blake, R. G. (1991). Stacking energies in DNA. *J. Biol. Chem.* **266**, 15160–15169.
Dujon, B., *et al*. (1994). Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.
Kenyon, C. J. & Walker, G. C. (1980). DNA-damaging agents stimulate gene expression at specific loci in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **77**, 2819–2823.
Kowalski, D. & Eddy, M. J. (1989). The DNA unwinding element: a novel, *cis*-acting component that facilitates opening of the *Escherichia coli* replication origin. *EMBO J.* **8**, 4335–4344.
Kowalski, D., Natale, D. A. & Eddy, M. J. (1988). Stable DNA unwinding, not breathing, accounts for single-strand-specific nuclease hypersensitivity of specific A + T-rich sequences. *Proc. Natl Acad. Sci. USA*, **85**, 9464–9468.
Little, J. W. & Mount, D. W. (1982). The SOS regulatory system of *Escherichia coli*. *Cell*, **29**, 11–22.
Liu, L. F. & Wang, J. C. (1987). Supercoiling of the DNA template during RNA transcription. *Proc. Natl Acad. Sci. USA*, **84**, 7024–7027.
Malkhosyan, S. R., Panchenko, Y. A. & Rekesh, A. N. (1991). A physiological role for DNA supercoiling in the anaerobic regulation of colicin gene expression. *Mol. Gen. Genet.* **225**, 342–345.
Natale, D. A., Schubert, A. E. & Kowalski, D. (1992). DNA helical stability accounts for mutational defects in a yeast replication origin. *Proc. Natl Acad. Sci. USA*, **89**, 2654–2658.
Sharp, P. M. & Cowe, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast*, **7**, 657–678.
Tsao, Y.-P., Wu, H.-Y. & Liu, L. F. (1989). Transcription-driven supercoiling of DNA: direct biochemical evidence from *in vitro* studies. *Cell*, **56**, 111–118.
Umek, R. M. & Kowalski, D. (1990). The DNA unwinding element in a yeast replication origin functions independently of easily unwound sequences present elsewhere on a plasmid. *Nucl. Acids Res.* **18**, 6601–6605.
Yamamoto, N. & Droffner, M. L. (1985). Mechanisms determining aerobic or anaerobic growth in the facultative anaerobe *Salmonella typhimurium*. *Proc. Natl Acad. Sci. USA*, **82**, 2077–2081.
Zaret, K. S. & Sherman, F. (1982). DNA sequence required for efficient transcription termination in yeast. *Cell*, **28**, 563–573.

***Edited by P. E. Wright***

*(Received 27 July 1995; accepted 11 October 1995)*