Pacific Symposium on Biocomputing 6:103-114 (2001)

STRESS-INDUCED DNA DUPLEX DESTABILIZATION IN TRANSCRIPTIONAL INITIATION

C.J. BENHAM
Department of Biomathematical Sciences, Box 1023, Mount Sinai School of Medicine, 1 Gustave Levy Place, NY, NY 10029, USA

Stress-induced destabilization of the DNA double helix (SIDD) is involved in several mechanisms by which transcription is regulated. This paper describes a computational method for predicting the locations and extents of destabilization as functions of DNA sequence and imposed superhelical stress. This method is used to investigate several transcriptional regulatory events. These include IHF-mediated activation of gene expression in *E. coli*, the bimodal control of the initiation of transcription from the human *c-myc* gene, and the determination of the minimal requirements for transcriptional activity in yeast. Collaborations with experimental groups have established the central role of SIDD in each of these processes.

1 Introduction

Initiation of transcription requires the two strands of the DNA double helix to transiently separate. This gives the polymerase complex access to the template bases of the encoding DNA, enabling construction of an RNA molecule having the complementary base sequence. So regulation of the initiation of gene expression requires stringent *in vivo* control of the locations and occasions of strand separation events.

The *in vivo* regulation of DNA strand separation commonly involves interactions with other molecules such as transcription factors, activators, inhibitors, and other DNA binding proteins. Several of these regulatory molecules must bind in a single strand-specific manner[1] that requires a pre-existing region of strand separation. The FBP protein, whose binding regulates *c-myc* oncogene transcription, is an example of this class.[2] However, even in cases where binding molecules actively participate in opening the DNA duplex, the site involved may need to be partially or entirely destabilized. One biologically important way in which the extent of DNA destabilization is regulated is through superhelical stresses imposed on the duplex.

The partitioning of chromosomal DNA into looped domains allows the linking numbers of individual domains to be independently regulated. (The linking number Lk of a domain is the total number of helical turns it contains when its central axis is planar.) When the linking number of a domain is smaller than the value Lk_0 characterizing its unstressed state, a negative linking difference $\alpha = Lk - Lk_0 < 0$ is imposed, and the domain is said to be (negatively) superhelical.

DNA superhelicity is stringently regulated *in vivo* by enzymatic activity.[3] It also is dynamically modulated during transcription, when the RNA polymerase pushes a bow wave of positive supercoils ahead of itself, and leaves a wake of negative supercoils behind.[4] In bacteria, the basal superhelicity is rapidly altered in response to environmental changes, including variations of anaerobicity [3] or osmolarity,[5] and induction of sporulation.[6] This rapidly changes the global patterns of gene expression.

When the imposed negative superhelicity is sufficient, it can destabilize the DNA duplex. A threshold stress level must be surpassed before significant destabilization occurs. Beyond that threshold, duplex destabilization is not uniformly distributed along the sequence, but rather is highly concentrated at a relatively small number of positions.[7,8,9] Local sites of strand separation, the most extreme form of duplex destabilization, can be induced by levels of negative superhelicity well within the physiological range.[8] Partial destabilizations also can occur, in which the the free energy needed to separate the duplex at a specific site is fractionally decreased by the imposed superhelicity.

Superhelicity globally couples together the conformations of all the base pairs within a domain. This coupling occurs because strand separation at any location, by changing the local helical twist, alters the level of torsional stress throughout the domain and thereby affects the transition behavior of every other base pair. The resulting behavior of a domain is determined by a complex interaction between the energetics of the transition, which vary with base sequence and environmental conditions, and the energetics of the superhelical deformations. Whether transition occurs at a given site depends not just on its local properties, such as its thermodynamic stability, but also on how that transition competes with all other transitions throughout the domain. Small changes in base sequence at one position can affect the probability of transition of sites thousands of base pairs away. Opening of one location can be coupled to the reversion back to B-form of remote opened sites.[10]

1.1 DNA Superhelicity Regulates Transcription

DNA superhelicity exerts a wide variety of effects on gene expression. The abundances of many *E. coli* proteins vary significantly with superhelicity.[11] The level of superhelicity at which transcription is optimized varies significantly from gene to gene. For example, although transcription from the *ampR*, *tetR* and *rep* genes of pBR322 all are enhanced by superhelicity, each is optimized at a different level of supercoiling.[12] In a more complex interaction, the *ilv*PG promoter is activated by integration host factor (IHF) binding, but only when the DNA is negatively supercoiled.[13] Conversely, negative superhelicity inhibits expression of GyrA, the A subunit of the gyrase protein that negatively supercoils DNA. This arrangement provides a mechanism for the homeostatic control of supercoiling.[14] The sensitivities of specific promoters to superhelicity enables the rapid changes of superhelicity that are induced in prokaryotes by alterations of environmental factors to quickly alter the global expression pattern of the entire chromosome.

One potentially important way in which changes of substrate superhelicity can alter transcriptional initiation rates is by affecting the energetics of open complex formation within the promoter. Processes that destabilize this opening site may increase transcription rates, while processes that stabilize it may decrease transcription rates.[15,16] These effects may either occur directly, or be modulated through the activities of transcription factors or other DNA binding proteins. The global coupling induced by superhelicity enables structural transitions occurring at one position within a domain to influence the conformations at other sites, and thereby exert regulatory effects over long distances. Examples include enhancers and silencers, which regulate gene expression in a distance- and orientation-independent manner, even at kilobase separations from the promoters they influence. Such processes enable a single regulatory region to simultaneously affect the activities of multiple promoters.

The essential role that local stress-induced duplex destabilization (SIDD) plays in transcription is emphasized by recent work defining the minimal transcriptionally active system in yeast.[17] There it was shown that transcription from the yeast Cup1 gene required *only* RNA polymerase and negative superhelicity - no other factors or molecules were necessary. Because the Cup1 gene is thought to be archaic, this suggests that primordial transcription originally may not have required ancillary molecules. These may have evolved later to exert more sophisticated levels of both positive and negative control.

2 The Computational Analysis of Superhelical Destabilization

At thermodynamic equilibrium a population of identical molecules is distributed among its available states, with the fractional occupancies of states decreasing exponentially as their free energies increase. Indexing the states by i and denoting the free energy of state i by G_i, the fractional occupancy of state i at equilibrium is

$$p_i = \frac{e^{(-G_i/RT)}}{Z}, \tag{1}$$

where R is the gas constant and T is the absolute temperature. (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.) As these probabilities must sum to unity, the normalizing factor Z - also called the partition function - is

$$Z = \sum_i \exp^{(-G_i/RT)} .$$

From these expressions the ensemble average value of any parameter at equilibrium may be calculated.

In our situation a linking difference α is imposed on a DNA molecule containing N base pairs of specified sequence. This topological condition can be accommodated by many combinations of structural deformations and conformational transitions. Here we model α as being partitioned among three factors. We designate the states of strand separation by defining N binary variables $n_j,\ j = 1, ..., N$ with $n_j = 1$ when base pair j is separated and $n_j = 0$ otherwise. A specific state of base pairing is identified by specifying the values of all these n_j's. The total number of open base pairs in this state is $n = \sum n_j$, and they occur in r runs. This transition decreases the total twist of the domain by n/A turns, where $A = 10.4$ base pairs per turn is the helicity of unstressed B-form DNA.[18] Torsional stresses will remain unless n has the precise value $n = -\alpha A$ that exactly relaxes the domain. Because single strands of DNA are much more flexible than is the B-form duplex,[19] the separated strands in a denatured region will tend to twist around each other in response to these 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 the above two deformations. (We need not decompose α_r further since the free energy associated to it is known from experiments.) The superhelical constraint couples these three deformations *via* the conservation equation

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

The free energy associated to a state contains contributions from each of these three factors. The free energy G_s needed to separate n base pairs in r runs is

$$G_s = ar + \sum_{j=1}^{N} b_j n_j$$

The values of $b_j,\ 1 \leq j \leq N$, are the energies needed to denature the base pair at each position j. This energy depends on the identity of the base pair itself, and on the identities of its neighbors. It is the sequence dependence of this energy that causes destabilization patterns to be non-uniform. Both the entropic and enthalpic components of b_j have been measured to high accuracy for all ten types of neighbor base pairs under a range of environmental conditions.[20] Also, a is the free energy required to initiate a run of separation; a has been measured experimentally to lie between 10 and 12 kcal/mol, depending on environmental conditions.[9,21,22]

The total twist $\mathcal{T}$ associated with the separated regions is

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

where τ_j is the local helicity (radians per base pair length) at each separated position j. A Hooke's Law free energy is associated to this deformation

$$G_t = \frac{C}{2} \sum_{j=1}^{N} n_j \tau_j^2.$$

The effective torsional stiffness C associated with this deformation has been measured experimentally.[9,22]

Lastly, the free energy G_r associated with residual linking α_r has been determined experimentally to be quadratic in α_r to high accuracy[23]:

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

K has been measured at various temperatures and ionic strengths.[9,22,24,25]

The total free energy associated to a state is the sum of these three contributions:

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

Because the free energies associated with each type of deformation have been experimentally determined, often using several different techniques under a range of conditions, there are **no free parameters** in any of the analyses reported below.

The partition function Z governing this system is: [26]

$$Z = \sum_{n_1=0}^{1} \cdots \sum_{n_N=0}^{1} \left\{ Q(n) \exp\left(-\beta \sum_{j=1}^{N} \Big\{(a + b_j)n_j - a n_j n_{j+1}\Big\} \right) \right\}, \qquad (2)$$

where

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

Three theoretical techniques have been developed by this research group to analyze superhelically driven structural transitions in DNA - an exact method,[26] a Monte Carlo sampling method,[27] and an approximate method.[9,28] The most flexible and efficient of these is a new generalization of the approximate method.

Here the discrete states of strand separation are ordered according to the size of their contributions to the partition function, largest to smallest. This is equivalent to ordering them by energy, smallest to largest. The state having minimum free energy G_{min} is determined, then an energy threshold θ is specified, and all states i are found whose free energies exceed G_{min} by no more than this threshold amount. An approximate partition function Z_{cal} is computed from this collection of low energy states to be

$$Z_{cal} = \sum_{i|G(i)-G_{min}<\theta} e^{-G(i)/RT}.$$

Approximate ensemble average (i.e. equilibrium) values are computed for all parameters of interest.

Although high-energy states are *individually* exponentially less populated than low energy states at equilibrium, they are so numerous that their *cumulative* contribution to the equilibrium still may be significant. So the next step is to estimate the aggregate influence of the states that do not satisfy the threshold condition. Originally this was done by a density of states procedure.[9,28] Now the exact method is used to evaluate precisely how the accuracy of these approximate calculations depends on the value of the threshold θ.[26]

This approximate method scales approximately quadratically with molecular length, although the details of the base sequence can exert important modulating effects. There is no simple scaling law with imposed superhelicity α. When α is either slightly negative or very negative the number of states satisfying a threshold condition can be relatively small. But in the intermediate range it can be much larger, resulting in a slower calculation. In practice, high accuracy is achieved using moderate values of θ (viz. more than four significant digits of accuracy in all parameters when $\theta = 12$ kcal/mol. at $T = 310$K). The calculations implementing the approximate method are reasonably efficient, often requiring less than 5 CPU minutes to analyze a 5kb sequence on an R10000 processor at physiological superhelicities. This efficiency makes the approximate method the technique of choice under most circumstances.

The biologically most important information involves the locations of either separated or destabilized sites, and the extents of their disruptions. For the first, we calculate the ensemble average probability $p(x)$ of separation of the base pair at each position x along the sequence. The graph of $p(x)$ *vs* x, called the transition profile, displays the regions of the sequence that have significant probabilities of opening. (Figure 2 below shows how the transition probabilities of a region in the *c-myc* gene regulatory element vary with superhelicity.)

A more sensitive measure of destabilization is given by the incremental free energy $G(x)$ needed to separate the base pair at position x[10,29]. This quantity is calculated as

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

where $\bar{G}$ is the ensemble average value of the free energy G, and $\bar{G}(x)$ is the average of that parameter over all states in which the base pair at position x is denatured. A value of $G(x)$ near or below zero indicates an essentially completely destabilized base pair, while positive values of $G(x)$ occur for base pairs where incremental free energy is needed to assure separation. Regions of partial destabilization are indicated by intermediate $G(x)$ values. Stress-induced duplex destabilization (SIDD) profiles, plots of $G(x)$ *vs* x, show regions of the sequence where superhelical stresses destabilize the duplex. Destabilization is usually confined to discrete sites, while most of the sequence experiences essentially no destabilization.

SIDD profiles are more informative than transition profiles because they also depict sites where the amount of free energy needed to induce separation is fractionally decreased. This will be important when duplex opening occurs by processes that can provide sufficient free energy to cause local separation only if the DNA site involved already is marginally destabilized. Such regions could be biologically important as sites which stresses render vulnerable to opening by enzymatic or other processes.

2.1 Tests of Accuracy Against Experimental Measurements

Superhelical DNA may be experimentally probed for unpaired regions using small molecules, including $KMnO_4$. Alternatively, DNA may be treated with single strand-specific endonucleases, enzymes that cut single stranded DNA but not the double helix. Following nuclease digestion the sites of enzymatic cleavage can be found by sequencing. Both small molecules and nucleases are used to probe for unpaired regions *in vivo*.[30,31]

Sample calculations have been performed on several DNA molecules for which experimental data on the locations and extents of superhelical denaturation are available.[9] In all cases the results of these calculations are in precise quantitative agreement with experimental measurements. The sites of separation were exactly predicted, and the calculated amounts of opening at each site agreed precisely with experimental measurements. The linking differences predicted to drive specific amounts of separation are within one turn of their observed values over the whole range of superhelicities where experiments were performed. This reflects the limit of accuracy with which extents of transition can be experimentally measured. And most importantly, the major changes in the locations of separated regions that result from minor sequence alterations were precisely predicted. This high accuracy has been achieved - *without free parameters* - in the analysis of every superhelical DNA molecule for which experimental data has become available, both *in vitro* and *in vivo*.[2,9,10,26,30]

These results show that this investigator's analytic methods produce quantitatively accurate predictions of even the fine details of the strand separation transition in superhelical DNA molecules. This justifies their use to predict the superhelical destabilization behavior of other molecules, on which experiments have not been performed. These calculations assume a linking difference that correspond to a moderate physiological superhelix density of $\sigma = \Delta Lk / Lk_o = -0.055$.

3 SIDD in Transcriptional Regulation

I. IHF-Mediated Gene Activation in Bacteria The *ilv*PG promoter of *E. coli* is activated by the binding of integration host factor (IHF) at a position 100 bp upstream from the transcription start site. This binding-induced activation only occurs when the DNA is negatively superhelical, not when it is relaxed. SIDD analysis of an experimental plasmid used to investigate this regulation showed that negative superhelicity strongly destabilizes the DNA duplex at the IHF binding site, as shown in Figure 1. This suggests that destabilization might be involved in the mechanism of superhelically induced IHF activation. It was proposed that IHF binding forces this region back into B-form, which transfers the superhelical destabilization to the next most easily destabilized site, located in the -10 region of the promoter, thus enhancing the rate of transcriptional initiation.

This mechanism for activation by the transmission of stress-induced destabilization was quickly experimentally proven.[32,33] In the absence of IHF, $KMnO_4$

binding showed the predicted SIDD site to be open and the -10 region of the promoter to be closed in superhelical plasmids. In the presence of IHF the situation was reversed: the SIDD site at the IHF binding position was closed, and the promoter was open.

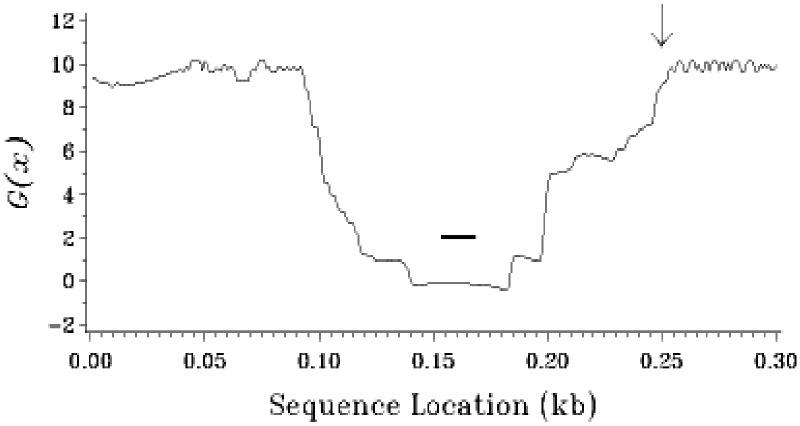

Figure 1: The SIDD profile of the upstream regulatory region of the *ilv*PG promoter is shown. The IHF binding site is indicated by a bar, and the transcription start site is shown by an arrow. Transcription proceeds to the right.

A computational search is being performed to find all *E. coli* ORFs with the attributes needed to be regulated in the same manner. All locations in the *E. coli* genome are found where a strong SIDD site coincides with a strong IHF binding site, and is located in a non-coding region upstream from an ORF. Our preliminary analysis has found 125 ORFs with this arrangement. These predictions are being tested using expression arrays. This is the first strategy to investigate a global genomic regulatory mechanism.

II. Regulation of the *c-myc* **Oncogene** The protein encoded by the *c-myc* gene is involved in cell growth, proliferation, differentiation and apoptosis. Its cellular abundance is regulated primarily at the transcriptional level, so dysregulation of the *c-myc* gene causes a variety of problems. Stable four-fold over-expression is oncogenic.[34] Indeed, even a transient pulse of high expression induces tumor growth and genomic instability.[35] Conversely, a decrease in *c-myc* expression to half its basal level prolongs the cell cycle.[36] So proper control of cell division requires that *c-myc* transcription be maintained within strict limits.

$KMnO_4$ binding shows the presence *in vivo* of a denatured site in the upstream control region of the *c-myc* oncogene.[31] The location of this so-called FUSE element exactly coincides with the strongest SIDD site in the 5' upstream flank of this gene, as shown in Figure 2. This site has a complex pattern of opening. As the negative superhelicity becomes more extreme, the first site to open is the downstream, promoter-proximal part of FUSE. After this opening is essentially complete, further superhelicity drives the opening of the promoter-distal region. So the strand opening of the FUSE element occurs in a bipartite manner.

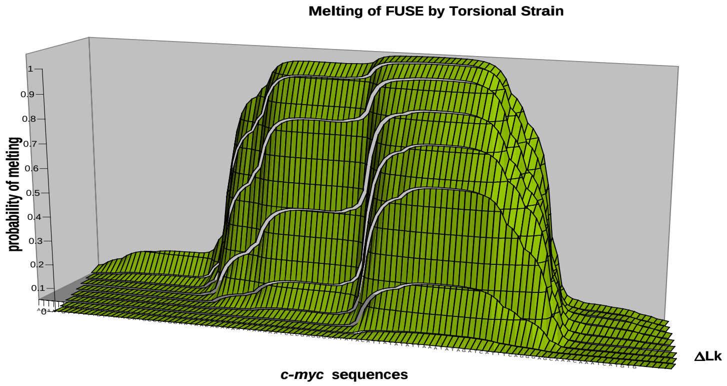

Figure 2: The transition probabilities of the region containing the FUSE element upstream of the *c-myc* gene are shown as a function of superhelicity. The early melting portion which binds to the activating domain of FBP is th the right, and the late melting portion which binds to the repressor domain of FBP is to the left.

Subsequent experiments have shown that the initiation of transcription from the *c-myc* oncogene is regulated by binding of the FBP protein to the single stranded FUSE element. FBP has both an activator domain and a repressor domain. These exert their respective effects on transcription when they bind to distinct but contiguous sites within the FUSE element. (The repressor domain is dominant: When both domains are bound the effect is repression.) The sites where the activator and repressor domains bind are precisely the early-opening and late-opening portions of FUSE, respectively. In each case binding requires the site involved to be single stranded. In this way both activation and repression of *c-myc* transcription by FBP are regulated by superhelical destabilization.[2]

III. Minimum Requirements for Transcription This investigator has collaborated with David Clark to characterize the minimal system required for the initiation of transcription from the *pCUP1* promoter of yeast.[17] We found that the only requirements were RNA polymerase and a substrate DNA that was sufficiently negatively supercoiled to permit destabilization near the promoter. Perhaps surprisingly, no other molecules were required for activity. Indeed, the site of destabilization need not precisely coincide with the transcription start site. Apparently the polymerase can use any destabilized location to enter the duplex. Once in, it can find the correct start position by a random search.

4 Discussion

The work reported here shows that SIDD calculations provide highly precise predictions of the locations of stress-destabilized sites in DNA sequences that can illuminate a variety of transcriptional regulatory processes. Although the conditions assumed in these calculations are much simpler than those prevailing *in vivo*, the sites of predicted destabilization agree precisely with the actual *in vivo* strand separation behavior in all cases examined to date. As shown in the case of *c-myc* regulation, even the fine details of strand opening that govern both its transcriptional activation and repression are accurately predicted.

These results demonstrate that structural properties of DNA, and specifically stress-regulated destabilization, are essential participants in the mechanisms by which specific transcriptional events are controlled. The calculation of SIDD profiles also is providing crucial new insights into the modes of activity of several other regulatory mechanisms, in addition to those described here.

Acknowledgments

The work described here was supported by grants DBI 99-04549 from the National Science Foundation, and RO1-GM47012 and RO1-HG01973 from the National Institutes of Health.

References

1. L. Rothman-Denes *et al*, *Cold Spring Harbor Symp. Quant. Biol.* **63**, 63 (1999).
2. L. He *et al*, *EMBO J.* **19**, 1034 (2000).
3. L.-S. Hsieh, R. Burger, and K. Drlica, *J. Mol. Biol.* **219**, 443 (1991).
4. L. Liu and J.C. Wang, *Proc. Nat'l. Acad. Sci. USA* **84**, 7024 (1987).
5. C. Higgins *et al*, *Cell* **52**, 569 (1988).
6. W. Nicholson and P. Setlow, *J. Bacterio.* **172**, 7 (1990).
7. C.J. Benham, *Proc. Nat'l. Acad. Sci. USA* **76**, 3870 (1979).

8. D. Kowalski, D. Natale, and M. Eddy, *Proc. Nat'l. Acad. Sci. USA* **85**, 9464 (1988).
9. C.J. Benham, *J. Mol. Biol.* **225**, 835 (1992).
10. C.J. Benham, *J. Mol. Biol.* **255**, 425 (1996).
11. T. Steck *et al*, *Molec. Micro.* **10**, 473 (1993).
12. J. Brahms *et al*, *J. Mol. Biol.* **181**, 455 (1985).
13. S. Parekh, S. Sheridan and G.W. Hatfield, *J. Biol. Chem.* **271**, 20258 (1996).
14. R. Menzel and M. Gellert, *Cell* **34**, 105 (1983).
15. H. Drew, J. Weeks, and A. Travers, *EMBO J.* **4**, 1025 (1985).
16. J.C. Wang and A. Lynch, *Curr. Opin. Genet. Dev.* **3**, 764 (1993).
17. B. Leblanc, C.J. Benham, and D.J. Clark, *Proc. Nat'l. Acad. Sci. USA* **97**, to appear (2000).
18. J.C. Wang, *Proc. Nat'l. Acad. Sci. USA* **76**, 200 (1979).
19. V. Bloomfield, D. Crothers, and I. Tinoco in*Physical Chemistry of Nucleic Acids*, (Harper and Row, New York, 1974).
20. G. Steger, *Nucl. Acids Res.* **22**, 2760 (1994).
21. B. Amirikyan, A. Vologodskii, and Y. Lyubchenko, *Nucl. Acids Res.* **9**, 5469 (1981).
22. W.R. Bauer and C.J. Benham, *J. Mol. Biol.* **234**, 1184 (1993).
23. W.R. Bauer and J. Vinograd, *J. Mol. Biol.* **47**, 419 (1970).
24. R. Depew and J.C. Wang, *Proc. Nat'l. Acad. Sci. USA* **72**, 4275 (1975).
25. D. Pulleyblank *et al*, *Proc. Nat'l. Acad. Sci. USA* **72**, 4280 (1975).
26. R. M. Fye and C. J. Benham, *Phys. Rev. E* **59**, 3408 (1999).
27. H. Sun *et al*, *J. Chem. Phys.* **103**, 8653 (1995).
28. C.J. Benham, *J. Chem. Phys.* **92**, 6294 (1990).
29. C.J. Benham, *Proc. Nat'l. Acad. Sci. USA* **90**, 2999 (1993).
30. A. Aranda *et al*, *Yeast* **13**, 313 (1997).
31. G. Michelotti *et al*, *Molec. Cell Biol.* **16**, 2656 (1996).
32. S. D. Sheridan, C. J. Benham, and G. W. Hatfield, *J. Biol. Chem.* **273**, 21298 (1998).
33. S. D. Sheridan, C. J. Benham, and G. W. Hatfield, *J. Biol. Chem.* **274**, 8169 (1999).
34. D. Aghib *et al*, *Oncogene* **6**, 707 (1990).
35. D. Felsher and J. Bishop, *Proc. Nat'l. Acad. Sci. USA* **96**, 3940 (1999).
36. M. Shichiri, K. Hanson, and J. Sedivy, *Cell Growth Differ.* **4**, 93 (1993).