

The Approximate Algorithm for Analysis of the Strand Separation Transition in
Superhelical DNA Using Nearest Neighbor Energetics

Chengpeng Bi
UC Davis Genome Center
University of California
One Shields Avenue
Davis, CA 95616
cbi@ucdavis.edu

Craig J. Benham
UC Davis Genome Center
University of California
One Shields Avenue
Davis, CA 95616
cjbenham@ucdavis.edu

Keywords: DNA, Stress-induced duplex destabilization (SIDDD), nearest neighbor energetics, superhelicity

Abstract

Accurate methods to computationally search genomic DNA sequences for regulatory regions have been difficult to develop. Conventional string-based methods have not been successful because many types of regulatory regions do not have recognizable motifs. And even when a sequence pattern is known to be associated with a class of regulatory regions it commonly is necessary, but not sufficient, for function. This suggests that other attributes, not necessarily strongly correlated with the details of base sequence, are involved in regulation. Here we present a computational method to analyze the propensity of superhelically stressed DNA to undergo strand separation events, as is required for the initiation of both transcription and replication. We build *in silico* models to analyze the statistical mechanical equilibrium distribution of a population of identical, stressed DNA molecules among its states of strand separation. In this phenomenon, which we call stress induced duplex destabilization (SIDDD), a state energy is determined by the energy cost of opening the specific separated base pairs in that state, and the energy relief from the relaxation of stress this affords. We use experimentally measured values of all energy parameters, including the nearest neighbor energetics known to govern DNA base pair stability. We perform a statistical mechanical analysis in which the approximate equilibrium distribution is calculated from all states whose free energies do not exceed a user-defined threshold. This provides the most general and efficient computational approach to the analysis of this phenomenon. The algorithm is implemented in C++, and its performance is analyzed.

INTRODUCTION

It has proven to be very difficult to computationally identify DNA regulatory regions using the conventional bioinformatics methods of genomic sequence analysis. This is because some sites, such as yeast transcription termination regions, do not have identifiable consensus sequences or motifs. And for those motifs that are known to have consensus sequences (such as AATAAA positioned 30 bp upstream from a polyadenylation site in higher eukaryotes) the presence of the motif is necessary, but not sufficient, for function. So any search based on such a motif will have unacceptably high false positive rates. For this reason we have focused on building computational methods to predict regulatory sites based on structural attributes of stressed DNA.

The B-form DNA structure is not invariant. Local DNA strand separation is a necessary step in the initiation of transcription and of replication, and also is involved in other processes. For this reason the locations and extents of strand separation must be stringently controlled *in vivo*. A closed circular DNA molecule is topologically constrained by the constancy of its linking number Lk . All conformational rearrangements that do not break DNA strands must preserve this constraint. In a living cell the DNA is partitioned into topological domains, typically several kilobases in length, consisting either of individual circular molecules or of closed loops within a chromosome. These loop domains are formed by periodic attachments of the chromatin fiber to the nuclear matrix (Alberts et al., 2002). The topological constraint on a closed-loop domain is precisely equivalent to that on a circular molecule; in both cases the linking number Lk is fixed. This value is regulated *in vivo* by a variety of processes involving transient strand breakage and religation. In this way the actual linking number Lk can be varied from its relaxed value Lk_o , so a linking difference $\alpha = Lk - Lk_o$ is imposed. This phenomenon is called DNA superhelicity. As superhelicity is commonly negative *in vivo* (i.e. $Lk < Lk_o$ so $\alpha < 0$) it imposes untwisting torsional stresses on the DNA, which can destabilize the double helix at specific sites, a phenomenon which we call stress-induced duplex destabilization (SIDDD).

DNA superhelicity, which is closely regulated *in vivo*, can induce the formation of locally unpaired regions at defined locations within DNA domains. 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 α -lactamase gene, and the promoter region of the same gene (Benham, 1993; Kowalski et al., 1988). The initiation of replication in both prokaryotes and yeast has been shown to require the presence at a precise position of a site that is susceptible to superhelical strand separation (Kowalski and Eddy, 1989; Huang and Kowalski, 1993). When the base sequence of this site is altered, replication occurs *in vivo* only if the susceptibility to stress-induced denaturation at the correct position is retained. SIDDD sites have also been shown to occur at chromosomal attachment regions (Benham et al., 1997). These attachments are known to augment transcription, and to form barriers between independently regulated domains. Sites susceptible to DNA duplex destabilization also occur at binding sites for other

molecules such as transcription factors and other regulators. In several cases the regulatory proteins require locally denatured DNA to bind (Rothman-Denes et al., 1998).

The approximate SIDD analysis method that was originally developed used copolymeric energetics, in which one value of the denaturation energy is assigned to each AT base pair, and another value to every GC base pair. However, it is known that the thermodynamic stability of DNA is significantly modified by the near neighbor identities. Moreover, specific base pairs can be modified in ways that alter their stability. Examples include base methylation, formation of adducts, ligand binding and the presence of abasic sites. In this paper we present the first computational method for analyzing SIDD in which the separation energies governing each base pair can be individually assigned. We have implemented this approach in an efficient approximate statistical mechanical algorithm, and assessed its operational characteristics.

NEAREST NEIGHBOR ENERGETICS OF LOCAL DENATURATION

Consider an N base-pair circular DNA molecule whose base sequence is $p_1p_2\dots p_j\dots p_N$, where p_i is either A, T, G or C. Suppose negative superhelicity is imposed with linking difference fixed at α . There are 2^N states of strand separation possible for this molecule. If a given state has n denatured base pairs under this constraint, the residual linking difference is determined as

$$\alpha_r = \alpha + n/A - T \quad (1)$$

Here A is the helical repeat, 10.4 bases per turn (Wang, 1979), and T is the total twist of the open regions. If the helical twist of each open base pair is τ rad/bp, then the total twist T is,

$$T = n\tau / 2\pi \quad (2)$$

The free energy G associated to each such state is comprised of three terms: the chemical energy (G_c) for separation of strands, the torsional energy (G_τ) for rotation of the single strands within denatured regions, and the residual supercoiling free energy (G_{res}). The chemical energy is associated with the set of base pairs that are open in the given state. Its values may be assigned individually. In the sample calculations reported here we use the experimentally determined values of the near neighbor energetics. The formulae for calculating these expressions are given below.

The chemical free energy of denaturation (strand separation), G_c , is divided into two contributions: the initiation free energy needed to nucleate a run of transition (a) and the incremental free energy of separating each base-pair in a denatured region (b).

$$G_c = ar + \sum_{i=1}^r \sum_{j=1}^{n_i} b_j \quad (3)$$

Here r is the number of runs. A run is a region composed entirely of separated base pairs. For the approximate method, $r \leq 3$. The value $a \approx 10.16$ kcal/mol was obtained by fitting to the experimental data (Kowalski et al., 1988) using nearest neighbor energetics.

The quantity b_j in equation (3) is the free energy needed to separate the j th base-pair in run i and can be computed using nearest neighbor energetics,

$$b_j = 0.5(\Delta G(p_{j+1}, p_j) + \Delta G(p_{j-1}, p_j)) \quad (4)$$

Here $\Delta G(p_{j+1}, p_j)$ is the right-side neighbor free energy and $\Delta G(p_{j-1}, p_j)$ is the left-side free energy. The p_j is the base at j th position. For a closed circular DNA with length of N , wrapping around is assumed: if $j = 0$, then $j \leftarrow N$; if $j > N$, then $j \leftarrow j - N$. A neighbor free energy, $\Delta G(p_i, p_j)$, is calculated according to the thermodynamic data including the neighbor enthalpy, ΔH and entropy, ΔS measured by Klump (Steger, 1994) and the absolute temperature T ,

$$\Delta G(p_i, p_j) = \Delta H(p_i, p_j) - T\Delta S(p_i, p_j) \quad (5)$$

Here the (p_i, p_j) is a neighbor base-pair, any combination in the base neighborhood space $\{A, T, C, G\} \times \{A, T, C, G\}$. The entropy also varies with the ionic concentration, given the experimentally determined entropy $\Delta S(p_i, p_j)$ in ionic concentration c_1 and the new entropy with ionic c_2 , $\Delta S'(p_i, p_j)$, can be deduced based on the thermodynamic principle,

$$\Delta S'(p_i, p_j) = \frac{\Delta H(p_i, p_j)}{16.6 \log(c_2 / c_1) + \frac{\Delta H(p_i, p_j)}{\Delta S(p_i, p_j)}} \quad (6)$$

The coefficient (16.6) was experimentally determined (Benham, 1992).

If there are n separated base-pairs in the denatured regions, the torsional free energy (G_τ) for rotation of the single strands within the regions is computed as,

$$G_\tau = Cn\tau^2 / 2 \quad (7)$$

where the constant C is the value of the torsional stiffness.

The free energy, G_{res} , associated with superhelical deformations has been measured by experimental techniques to be quadratic for the linking difference without strand separation and the residual linking difference with denaturation as well. The formula is expressed as,

$$G_{res} = \frac{K\alpha_r^2}{2} = \frac{K}{2} \left(\alpha + \frac{n}{A} - T \right)^2 \quad (8)$$

The coefficient K has been determined experimentally to vary inversely with the sequence length N , having the value $K \approx 2200RT/N$ at the physiological temperature $T = 310^\circ\text{K}$. R is the gas constant.

The total free energy for a state can be calculated as,

$$G = G_c + G_\tau + G_{res} = ar + \sum_{i=1}^r \sum_{j=1}^{n_i} b_j + Cn\tau^2 / 2 + \frac{K}{2} \left(\alpha + \frac{n}{A} - T \right)^2 \quad (9)$$

If we minimize the above equation with respect to τ , we have the relationship: $\alpha_\tau K = 2\pi C\tau$. Therefore equation (9) can be rewritten as,

$$G = ar + \sum_{i=1}^r \sum_{j=1}^{n_i} b_j + \frac{2\pi^2 CK}{4\pi^2 C + Kn} \left(\alpha + \frac{n}{A} \right)^2 \quad (10)$$

THE APPROXIMATE STATISTICAL MECHANICAL METHOD

A complete description of the statistical mechanical method for analyzing superhelical duplex destabilization has been presented elsewhere (Benham, 1990, 1992; Fye and Benham, 1999). Here we use an approximate approach that finds all states whose free energy does not exceed a specified threshold (Benham, 1990, 1992). The first step in this analysis is to determine the state of minimum free energy (G_{min}). Then a threshold energy value (θ) is specified, and all states (S) are searched whose free energy exceeds that of the minimum energy state by no more than the threshold θ . Expressions for the partition function and other important statistical mechanical quantities are evaluated from this collection of states, as described below. We calculate two equilibrium properties that describe the destabilization experienced by the input sequence at this stress level. First, the ensemble average probability $p(x)$ of the base-pair at each position x in the sequence is given by:

$$p(x) = \frac{\hat{Z}(x)}{\hat{Z}} \quad (11)$$

where \hat{Z} is the approximate partition function:

$$\hat{Z} = \sum_{s \in S} e^{-G_s / RT}, \quad S = \{s : G_s < G_{min} + \theta\} \quad (12)$$

$\hat{Z}(x)$ sums over all states X in which the base-pair at position x is open:

$$\hat{Z}(x) = \sum_{s(x) \in X} e^{-G_{s(x)} / RT}, \quad X = \{s(x) : G_{s(x)} < G_{min} + \theta\}, X \subset S \quad (13)$$

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 free energy $G(x)$ needed to separate the base-pair at position x (Benham, 1999). This quantity is calculated as:

$$G(x) = \bar{G}(x) - \bar{G} \quad (14)$$

where \bar{G} is the ensemble average free energy of the system:

$$\bar{G} = \frac{\sum_{s \in S} G_s \exp(-G_s / RT)}{\hat{Z}} \quad (15)$$

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

$$\bar{G}(x) = \frac{\sum_{s(x) \in X} G_{s(x)} \exp(-G_{s(x)} / RT)}{\hat{Z}(x)} \quad (16)$$

The plot of $G(x)$ versus x is called the (helix) destabilization profile.

ALGORITHM IMPLEMENTATION

The approximate algorithm converts the 2^N problem to its sub-problem by specifying an energy threshold (θ) and enumerates all those low energy states satisfying this criterion. The neglected high energy states can be estimated (Benham, 1990). An exact method was also developed (Fye and Benham, 1999), but it is very slow to execute owing to catastrophic cancellation and can be used to verify the approximate method.

Let W be the window size of a run. In practice $W \leq 250$ and the energy threshold $\theta = 12.0$ kcal/mol. Define the array $G_c[1:W, 1:N]$ storing all the denaturation energies for each window size and start position in sequence. Define the array $Y[1:W, 1:N]$ storing all the low free energies and $Y_e[1:W, 1:N]$ storing the Boltzmann's factors of Y by equation (12).

The algorithm 1 computes chemical energies for each window size and start position and then sorts them and find a minimum chemical energy among them. The first row of G_c holds all the minimum energy values for different windows. Here sorting is needed to speed up the searching of low energy states later on.

Algorithm 1: computing the chemical energy and sorting

input: window size W , sequence and thermodynamic data
output: sorted chemical energies $G_c[1:W, 1:N]$ and $minG_c$
for $i \leftarrow 1$ to W **do**
 for $j \leftarrow 0$ to N **do**
 compute $G_c[i, j]$ based on equations (3-6)
 sort on $G_c[i]$ // quick sort
for $i \leftarrow 1$ to W **do**
 $minG_c \leftarrow \min(minG_c, G_c[i, 1])$

The algorithm 2 computes the non-chemical energy part ($G_{nc}[1:W]$) comprising of torsional and residual free energy. This energy is only dependent on the window size for a given run.

Algorithm 2: computing the non- G_c free energy: $G_{nc}[1:W]$

input: α , W , C , K , and A

output: G_{nc} and $minG_{nc}$

for $n \leftarrow 1$ to W **do**

$$G_{nc}[n] = \frac{2\pi^2 CK}{4\pi^2 C + Kn} \left(\alpha + \frac{n}{A} \right)^2$$

$minG_{nc} \leftarrow \min(minG_{nc}, minG_{nc}[n])$

Algorithm 3: Find_States(1) // searching states w/ minimum energy in one-run

input: W , N , G_c , G_{nc}

output: minFlag and all the states with low energy

minFlag \leftarrow **false**

for $n \leftarrow 1$ to W **do**

for $j \leftarrow 0$ to N **do**

$E \leftarrow G_c[n, j] + G_{nc}[n]$

if $E < minG + \theta$ **then**

store(E , n , j) // store low energy state

if $E < minG$ **then**

$minG \leftarrow E$

minFlag \leftarrow **true**

else break

return minFlag

Algorithm 4: Find_States(2) // searching states w/ minimum energy in two-run

input: G_c , G_{nc} , $minG$, θ and W

output: minFlag and all the states with low energy

minFlag \leftarrow **false**

for each run pair(n_1 , n_2) & $n_1 + n_2 \leq W$ **do** // n_1 , n_2 are window sizes of two runs

if $G_c[n_1, 1] + minG_c + minG_{nc} \geq minG + \theta$ **then** next n_1

if $G_c[n_1, 1] + G_c[n_2, 1] + G_{nc}[n_1+n_2] \geq minG + \theta$ **then** next n_2

for $i \leftarrow 1$ to N **do**

if $G_c[n_1, i] + G_c[n_2, 1] + G_{nc}[n_1+n_2] \geq minG + \theta$ **then break**

for $j \leftarrow 1$ to N **do**

$E \leftarrow G_c[n_1, i] + G_c[n_2, j] + G_{nc}[n_1+n_2]$

if $E \geq minG + \theta$ **then break**

else

if two runs are not overlapped **then**

store(E , n_1 , i) and store(E , n_2 , j)

if $E < minG$ **then**

$minG \leftarrow E$

minFlag \leftarrow **true**

return minFlag

Define the function Find_States(runs) as the procedure to search all the low energy states and then store them. Depending on the number of runs, we call different functions. The algorithm 3 is such a function to search states with low energy in one-run states. For each state, a new minimum energy is recorded if found and set the flag in the variable minFlag. The function store will update the values Y and Y_e and compute the profiles based upon the equations (11-16). For each window, we only store the first position and accumulate the values: Y and product of Y_e and Y at that position. After the searching process was done, a filling procedure was applied and filling the downstream locations for each start position and window size. The calculation of probability and destabilization profiles is based on equations (11, 14).

The algorithm 4 does the same thing as algorithm 3 but with two runs. If number of runs is larger than one, we have to check if two runs are overlapped for each state. Two runs should have at least one base pairs gap to be considered as non-overlap.

For three runs, we apply the same idea as in algorithm 4. The main program will call the function Find_States for run 1, 2 and 3 in sequence. If a new minimum energy is found, then start over the searching process. The minimum value is most likely located in the one-run states. It is very rare to find a minimum in two-run and three-run states. The reason is the total energy increases by 10.16 kcal for each additional run. After the searching process ended, a filling procedure was applied and followed by computation of the profile values.

EVALUATION OF ENERGY PARAMETERS USING NEAREST NEIGHBOR ENERGETICS

Three energy parameters, the initial energy of a run separation (a), the torsional stiffness (C) and the quadratic coefficient (K), were determined by fitting to the experimental data (Kowalski et al., 1988). The experimental data used in this analysis are the locations and percentages of strand separation occurring in the pBR322 DNA molecule. The thermodynamic data (enthalpies and entropies) of nearest neighbor energy for base open are from Klump (Steger, 1994). Under the experimental conditions ($T = 310$ K, $\text{pH} = 7.0$, monovalent cation concentration = 0.01 M) three linking differences (-26, -28 and -32) were used, we can find the optimal values by fitting to the experimentally determined locations and extents of the destabilized sites. By numerical simulation we found that the initial energy (a) increases as the quadratic coefficient (K) increases. This is a linear relationship. On the other hand, the initial energy (a) is a logarithm function of the torsional stiffness (C). Based on previous experimental reports (Benham, 1992) we set the potential solution space as: $C = [1.5, 2.5]$, $K = [2100\text{RT}/N, 2500\text{RT}/N]$ and $a = [9, 11]$. By running the above algorithms, we found the best numerical solution is at $C = 1.91$ kcal per square radian, $K = 2200\text{RT}/N$ and $a = 10.16$ kcal per mol. The best fitness (RMS) is 0.056%. These estimated parameters are very close to the experimental measurements (Benham, 1992; Bauer and Benham, 1993).

ALGORITHM PERFORMANCE

There is a tradeoff between CPU time and accuracy. The more states found, the more time spent. There are 2^N states in total. The approximate method considers those states with low energy, a subset of the total. Appropriate setting of threshold and superhelical density can reach very good performance and high precision. In practice we set threshold 12.0 kcals and superhelical density -0.055. For the pBR322 plasmid DNA with length 4363 bp, it took about 20 seconds to complete on a 1.0 GHz PC machine. If the threshold is over 14.0 kcals, the CPU time will exponentially grow (Figure 1). The same situation holds when the superhelical density is less than -0.07. Under the physiological condition the superhelical density is around -0.06.

The memory space takes $O(N)$, N is the DNA length. In normal cases (threshold = 12.0 kcals and superhelical density = -0.055), the running time (t) is bounded by $O(N \log N) < t \leq O(N^2)$. In the worst case (higher negative superhelicity and/or higher threshold) running time is $O(N^3)$ and it takes over an hour to complete a 5 kb sequence. On normal cases, it takes around 30 seconds to finish a 5 kb DNA sequence on a 1.0 GHz PC.

DISCUSSION

In this paper we built *in silico* SIDD simulation models to predict those regulatory sites in DNA sequence. The nearest neighbor energetics was combined into the SIDD models and the approximate method was applied to analyze stress-induced duplex destabilization. The algorithms were implemented in C++ and the performance was efficient.

The profile results generated by this algorithm are similar to those by previous methods (Benham, 1999). Some slightly differences may imply more important usefulness in biological sense. This needs further development in the future.

The circular DNA can be analyzed by direct application of the methods described above. The linear DNA molecules are handled by adding a run of 50 consecutive GC nucleotide bases that conceptually closes them into circles (Benham, 1999). The modified linear DNA can be analyzed by the above methods.

The performance is also subject to the DNA length and composition. This algorithm was designed to compute DNA with length less than 10 kb. For longer or whole genomic DNA sequence, there is a windowing procedure to handle this (reported elsewhere). There is a web version of the algorithm which is accessible at <http://genomecenter.ucdavis.edu/sidd/>.

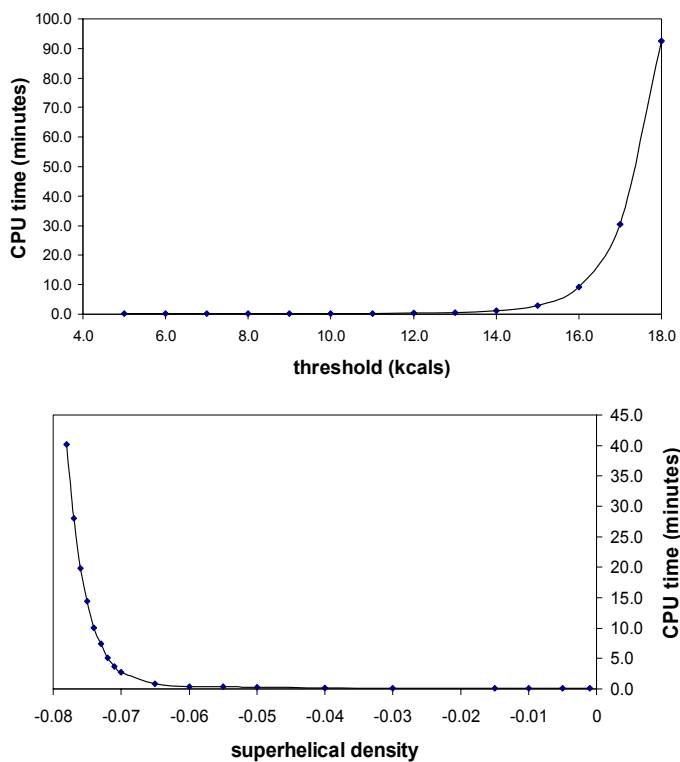


Figure 1. The CPU time varies with the threshold and superhelical density. Data were collected by running the program on a PC machine with 1.0 GHz Intel Pentium III and 512 MB memory. The compiler is GNU C++. The upper graph shows the relationship of time with threshold. After the threshold is over 14.0 kcal, the CPU time start to exponentially grow. The lower graph demonstrates the relationship of time with superhelical density. If the density is lower than -0.07 (high negative superhelicity), the CPU time grows exponentially with the absolute value of the density.

REFERENCES

1. Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*. Garland, NewYork.
2. Benham,C.J. (1990) Theoretical analysis of heteropolymeric transitions in superhelical DNA molecules of specified sequence. *J. Chem. Phys.*, **92**, 6294-6305.
3. Benham,C.J. (1992) Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, **225**, 835-847.
4. 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.
5. Benham,C.J. (1999) The Topological driven strand separation transition in DNA-methods of analysis and biological significance. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **47**, 173-198.
6. 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.
7. Breslauer,K., Frank,R., Bloecker,H. and Marky,L. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Nat'l. Acad. Sci. USA*, **83**, 3746-3750.
8. 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.
9. 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.
10. 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.
11. Kowalski,D., Natale,D. and Eddy,M. (1988) Stable DNA unwinding, not breathing, accounts for single-stranded specific nuclease hypersensitivity of specific A+T-rich regions. *Proc. Nat'l Acad. Sci. USA*, **85**, 9464-9468.
12. Rothman-Dees,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.
13. Steger,G. (1994) Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction. *Nucleic Acids Research*, **22**, 2760-2768.
14. Wang,J.C. (1979) Helical repeat of DNA in solution. *Proc. Nat'l. Acad. Sci. USA*, **76**, 200-203.