

WebSIDD – Computational Prediction of Stress-induced Duplex Destabilized Sites in Superhelical DNA

Chengpeng Bi and Craig J. Benham

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

ABSTRACT

This paper describes the WebSIDD program that predicts locations and extents of stress-induced duplex destabilization (SIDD) that occur in a double-stranded DNA molecule of specified base sequence, on which a specified level of superhelical stress has been imposed. The algorithm calculates the approximate equilibrium statistical mechanical distribution of a population of identical molecules among the accessible states. The input to the program is a DNA sequence, and its output is the calculated transition probability and destabilization energy of each base pair in the sequence. The structural and energy parameters used in the calculation are all determined experimentally, so there are no free parameters to be fit. Yet the output of the program is in quantitative agreement with experimental results in all cases where experiments have been performed. This method has illuminated the roles of SIDD properties in the regulation of diverse biological processes, including transcriptional initiation and termination, and the eukaryotic nuclear scaffold attachments that partition chromosomes into domains. The WebSIDD program is accessible at the web address: <http://genome.bme.ucdavis.edu/sidd/>. A detailed description of how to use this software is given in the text.

INTRODUCTION

DNA is constrained into topological domains *in vivo*, typically several kilobases in length, consisting either of circular molecule or of closed loops within a chromosome that are formed by periodic attachments of the chromatin fiber to the nuclear matrix (1). 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*, it imposes untwisting torsional stresses on the DNA which can destabilize the double helix.

The B-form DNA structure is not permanent. 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 occasions of strand separation must be stringently controlled *in vivo*. DNA superhelicity, which is closely regulated *in vivo*, can induce the formation of locally unpaired regions at defined sites within DNA molecules. 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 (2,3). 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 (4,5). 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. SIDD sites have also been shown to occur at chromosomal attachment regions (6). 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 (7).

Although mechanisms of DNA function involving stress-induced duplex destabilization (SIDD) commonly also involve complex interactions with other molecules, the intrinsic susceptibility to stress-induced base-unpairing at specific sites clearly is essential for activity. The local thermodynamic stability of DNA is low at these sites. However, this is a strictly local attribute of the sequence, depending on base pair identities, near neighbors, temperature and ionic strength (8-10). The susceptibility to SIDD is not determined by strictly local properties of DNA sequence, but rather by a global competition among all sites within a stressed domain. This is because the imposed stress couples together the behaviors of all base pairs that experience it. This occurs because opening any site alters its local helical twist, which changes the distribution of superhelicity and thereby affects the denaturation probabilities of every other base-pair in the domain. In this way superhelicity creates a global coupling among the conformational states of all base-pairs within a domain.

The free energy associated with negative DNA superhelicity in principle also could drive the formation of Z-DNA, cruciforms and H-DNA structures at susceptible sites. However, these structures, unlike local strand separation, are not known to play essential roles in biological processes. However, they can be analyzed using the same strategy as has been employed to calculate SIDD transitions.

In this paper we briefly describe the stress-induced DNA duplex destabilization (SIDD) model and the WebSIDD program we have implemented to predict the SIDD sites on a DNA molecule on which negative superhelicity has been imposed. The algorithm is based on a statistical mechanical SIDD analysis procedure that has been presented earlier (11,12). The SIDD properties can be computed based upon either copolymeric or near neighbor energetics. The input is a DNA sequence and the output is the transition probability profile and destabilization energy profile for each location of nucleotide bases in the molecule. A brief description of how to use the web software is given below. The WebSIDD software is web accessible at <http://genome.bme.ucdavis.edu/sidd/>.

ALGORITHM

A complete description of the statistical mechanical method for analyzing superhelical duplex destabilization has been presented elsewhere (11-13). Here we use an approximate approach that finds all states whose free energy does not exceed a specified threshold (2,11). The free energy G associated to each 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 each state of base-pairing of the DNA, and is computed based upon either copolymeric or near neighbor energetics. The formulae for calculating these expressions are given below.

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 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:

$$p(x) = \frac{\hat{Z}(x)}{\hat{Z}} \quad (1)$$

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

$$\hat{Z} = \sum_{s \in S} e^{-G_s/RT}, \quad S = \{s : G_s < G_{min} + \theta\} \quad (2)$$

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

$$\hat{Z}(x) = \sum_{s(x) \in X} e^{-G_{s(x)}/RT}, \quad X = \{s(x) : G_{s(x)} < G_{min} + \theta\}, X \subset S \quad (3)$$

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 (2). This quantity is calculated as:

$$G(x) = \bar{G}(x) - \bar{G} \quad (4)$$

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

$$\bar{G} = \frac{\sum_{s \in S} G_s \exp(-G_s / RT)}{\hat{Z}} \quad (5)$$

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

$$\bar{G}(x) = \frac{\sum_{s(x) \in X} G_{s(x)} \exp(-G_{s(x)} / RT)}{\hat{Z}(x)} \quad (6)$$

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

INPUT

The WebSIDD program has a user interface that enables the user to control the inputs through a single table (Figure 1). To trace the usage of the WebSIDD program, we ask each user to log on, providing their name and email address as options. There are two action buttons in the bottom of the table. The submit button activates the server program and performs the analysis, and the reset button resets all variables to their default values. Each parameter value has a bounded range, called its SIDD value range, chosen to correspond to biological possibilities, and a default value. This is done in order to prevent abnormal behaviors. This limit can be relaxed upon user's request. *If a field value is empty or not a number, the default value is set without warning.*

Types of DNA molecules. One can specify whether the sequence to be analyzed is to be regarded as circular or linear. These cases are handled differently, so one must be sure to check the right molecular topology. The default is circular DNA.

Types of energetics. There are two types of opening energies: either copolymeric or near neighbor. The copolymeric type assigns the same free energy value to every AT and a different value to every GC base pair. The near neighbor type assigns entropies and

enthalpies to each of the ten different neighbor types, as measured by Klump (10). The chemical free energy is calculated as $G_c = ar + \sum_{i=1}^N b_i$. Here b_i is evaluated according to Klump, r is the number of runs of denaturation, and a is the initial energy of nucleating a run of transition.

Temperature and salt concentration. The default values of temperature and salt concentration are 37°C (310 K) and 0.01 M respectively. All energy parameters are selected to be consistent with these default physiochemical conditions, as they are the experimental settings used in the mung bean nuclease digestion procedure by which stress-induced strand opening is most accurately assessed *in vitro* (3). However, we leave the possibility for the user to specify other conditions. The SIDD lower bound values for temperature and ionic strength are 237.0 K and 0.005 M, respectively; and upper bound values are 373.0 K and 1.0 M, respectively. We note, however, that the algorithm takes substantially longer to execute when a high temperature is selected.

The sperhelix density and threshold. The superhelical density (σ) is set to $\sigma = \alpha / Lk_0$, i.e. the linking difference divided by the link number Lk_0 of a relaxed B-form DNA. Given the DNA length N and a density σ , the linking difference is calculated as: $\alpha = \sigma N / A$. The helical repeat A is 10.4 bp per turn (14). The default value of superhelix density is $\sigma = -0.055$. The current SIDD version only allows negative superhelicity as occurs *in vivo* under normal physiological situations. The number of states satisfying any given threshold increases approximately exponentially with the absolute value of σ (12). However, at the default threshold energy (θ) of 12.0 kcal/mol, the program executes efficiently while providing high accuracy. As individual states become exponentially less frequently occupied as the threshold is increased, states beyond this threshold do not contribute significantly to the equilibrium. In practice this threshold should not be decreased below 10.0 kcal/mol, however, as this sacrifices accuracy. The limits on the WebSIDD value of θ are between 9.0 and 20.0 kcal/mol.

Open region size. In practice we set the maximum size (W) of the open region to be between 200 and 250 bp. In practice longer open regions do not occur under physiological conditions in molecules on the kilo-base length scale. The limits set in WebSIDD for this parameter W are between 50 and 250 bp. It is suggested that users do not change this value, as decreasing it could eliminate some low energy states and increasing it beyond the maximum length experienced will only slow the program without increasing its accuracy.

Quadratic coefficient (K). The coefficient, K , of the quadratic free energy is associated to the residual linking difference (α_r). The quadratic free energy is $G_{res} = K\alpha_r^2/2$. The value K was experimentally found to be $2220RT/N$ under moderate salt conditions (R is the gas constant, T is temperature and N is the DNA sequence length). Our simulation found the best value of K is $2200RT/N$ under 0.01 M salt concentration by using near neighbor energetics. The SIDD K value ranges from 2000 to 2400 RT/N.

Torsional stiffness (C). The torsional stiffness C is associated to the interstrand twisting of the two single strands comprising a locally denatured region. If the total number of denatured base pairs is n and each single strand twists about the other at rate τ (radians per base pair), then the torsional free energy is computed as: $G_\tau = Cn\tau^2/2$. The C value was found to increase as the ionic strength was lowered (9). Our simulation showed that the best-fitting C value of stiffness is 1.91 kcal/mol/rad². The WebSIDD program allows C value to range between 1.5 and 3.8 kcal/mol/rad².

Initial energy (a). The initiation free energy a is the energy needed to nucleate a run of transition. Its value is dependent on temperature and ionic strength. Under the default conditions $a = 10.16$ kcal/mol, but the WebSIDD program allows values between 9.0 and 11.0 kcal/mol.

DNA sequence. A user can either edit or copy and paste the DNA sequence to be analyzed into the appropriate text area in the WebSIDD interface window. The legal character set is {A, C, G, T, 0-9}. The program regards numerical input as line numbers,

and ignores the illegal characters. The allowed letters are not case-sensitive; either capitals or small letters will work. A sequence name is required for each new input. The default name is given as “seqtest1”. The sequence can be a natural or constructed plasmid, a whole viral sequence, or a DNA segment from a longer genome or chromosome. The length of the DNA sequence cannot exceed 10 kb.

Cutoff options. There are five levels for cutoff energy. Level 1 is 0.0 kcal/mol, Level 2 is 2.0, Level 3 is 4.0, Level 4 is 6.0 and Level 5 is 8.0 kcal/mol. The default value is level 4.

Profile options. The program provides a detailed profile including the base position, transition probability and destabilized free energy. The profile can be downloaded. Graphic profiles are also generated on-the-fly in GIF format. Given an energy cutoff, the program outputs possible locations and extents of SIDD sites.

Output control. There are four items in this multiple selection. One can choose from 0 to 4 items with any combination. The states item is checked as default. It will report number of states and the longest open regions found in each run of transition. By ticking the first two items, the input DNA sequence and parameter settings will be displayed in the output browser window. The ensemble quantities include (i) the ensemble average energy of the system (\bar{G}); (ii) the ensemble average number of open regions: $\bar{r} = \sum_{s \in S} r e^{-G_s / RT} / \hat{Z}$ ($r = 1, 2, 3$); (iii) the ensemble average number of open base pairs: $\bar{n} = \sum_{i=1}^N p(x_i)$; (iv) the ensemble total twists of open regions (\bar{T}) and (v) the ensemble average residual linking difference ($\bar{\alpha}_r = \alpha + \frac{\bar{n}}{A} - \bar{T}$).

OUTPUT

We use the standard experimental plasmid pBR322 to demonstrate the functionalities of the WebSIDD program. This plasmid has 4363 bps in a circular configuration. Its superhelical destabilization properties have been extensively studied, both experimentally and theoretically (3,11-13). We set all parameters at their default values, and checked all

items in the output control. We copied the pBR322 DNA sequence and pasted it into the text area of WebSIDD program, then clicked the “Submit” button to run the WebSIDD program on the server-side (One can obtain the example DNA sequence from the WebSIDD URL address: <http://genome.bme.ucdavis.edu/sidd/tutorial.html>). A separate browser window pops up in which the output will be displayed after the program completes, which in this case requires only a few seconds.

Profile output. If one chooses to output the profiles in text mode, one receives the output directly from the server. This contains four columns: the sequence location x , the nucleotide base, its transition probability $p(x)$, and destabilization free energy $G(x)$. The text profile can be downloaded. If one selects the graphics mode as well, graphs of the transition probability vs sequence and destabilization energy vs sequence are also generated on-the-fly. Figure 2 shows the calculated graphical profiles for pBR322 sequence. Given the cutoff level 4 (6.0 kcal/mol), a major destabilized (SIDD) and site (see Figure 2 in right graph) occurs at positions 3173-3317 and a minor SIDD site occurs at positions 4169-4332 (sites 2 and 3 combined). These results agree in detail with the sites of single-strand-specific endonuclease cleavage, found by Kowalski *et al.* (3). The major and minor SIDD sites occur at the terminator and the promoter regions of the *amp* gene, respectively (2).

DAN sequence and parameter settings. Our output also includes the whole input DNA sequence and the parameter settings. The parameter settings used in this demonstration are: Energetics: near neighbor; Molecule type: circular DNA; Threshold (θ): 12.00 kcals; Maximum open region (W): 200 bp; Temperature: 310.0 Kelvin Degree; Salt concentration: 0.010 M; Initial free energy (a): 10.16 kcals; Superhelix density (σ): -0.05500; Helical repeat (A) = 10.40 bp/turn; Torsional stiffness (C) = 1.91 kcals-bp/rad²; Quadratic coefficient (K): $2200RT/N$; Linking difference (α): -23.07356 turns (The sequence length (N) is 4363 bp).

State report. The state report generated by this example calculation is the following: The lowest free energy is 68.28628 kcals; Total states with low energy number 13543656. Of

these, 13852 contain one denatured region, 13529516 states contain two denatured regions, and 288 states contain three denatured regions. The longest open regions contain 127 bp in the one run states, 99 bp when there are two runs, and 44 bp when there are three open runs.

Ensemble quantities. There are five ensemble quantities calculated: The ensemble average free energy of the system is $\overline{G} = 69.83683$ kcal/mol. The average number of open regions is $\overline{r} = 1.07700$, and the average number of open base pairs is $\overline{n} = 66.10074$ bp. The average total twist of the open regions is $\overline{T} = -3.44458$ rad, and the average residual linking difference is $\overline{\alpha}_r = -13.27313$ turns.

DISCUSSION

In this paper we briefly describe the algorithm used by the WebSIDD program was based upon, and demonstrate how to use this program through our Web interface. To illustrate its use we performed a sample SIDD computation using the pBR322 DNA sequence. The important results of a SIDD computation are the transition probabilities and destabilization energies. The ranges of each user-specifiable input variable and output format also are described. These bounds can be relaxed upon special request from a user. In this implementation the sequence length is limited to 10 kbp. SIDD analysis of longer DNA sequences requires a windowing procedure that will be described elsewhere. The graphical output is intended to provide a quick overview of the results. Users are strongly recommended to save the text output of the profiles and plot the proper graphs as needed. Further questions concerning the SIDD computation can be directed to the authors. Supplemental materials including a tutorial can be found at the web address: http://genome.bme.ucdavis.edu/sidd/websidd_paper.htm.

ACKNOWLEDGEMENT

This work was supported in part by grants DBI 99-04549 from the National Science Foundation, and RO1-HG01973 from the National Institutes of Health, and by additional support from the Diversa Corporation.

REFERENCES

1. Alberts,B., Bray,D., Lewis,J., Roberts,K. and Watson, J. D. (1994) *Molecular Biology of the Cell*. Garland, NewYork.
2. 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.
3. 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.
4. 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.
5. 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.
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. 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.
8. 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.
9. Benham,C.J. (1992) Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, **225**, 835-847.

10. Steger,G. (1994) Thermal denaturation of double-stranded nucleic acids. *Nucleic Acids Res.*, **22**, 2760-2768.
11. Benham,C.J. (1990) Theoretical analysis of heteropolymeric transitions in superhelical DNA molecules of specified sequence. *J. Chem. Phys.*, **92**, 6294-6305.
12. Benham,C.J. (1992) The energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, **225**, 835-847.
13. 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.
14. Wang,J.C. (1979) Helical repeat of DNA in solution. *Proc. Nat'l. Acad. Sci. USA*, **76**, 200-203.

Figure Legends

Figure 1. User Interface to Advanced WebSIDD Computation. This is WebSIDD user interface showing the default parameter settings except setting the profile type as text and graphical, a sample DNA sequence and checking all items in output control. There are two action buttons: submit – to submit the job, reset – to set all the values to default states and clear the DNA sequence in the text area.

Figure 2. Profile output for pBR322 plasmid (4363 bp, circular DNA) showing the ensemble average probability of denaturation of each base pair (transitional profile) and the destabilized free energy of each base pair (destabilization profile); Given the cutoff energy 6.0 kcal/mol (level 4), the potential SIDD sites are also indicated below the graphs. Here we set open window size of 200 bp, the near neighbor energetics, temperature of 310 K, ionic strength of 0.01 M, superhelix density of -0.055, threshold energy of 12.0 kcals, $K = 2200RT/N$, $C = 1.91$ kcals per rad^2 and $a = 10.16$ kcals.

Figure 1

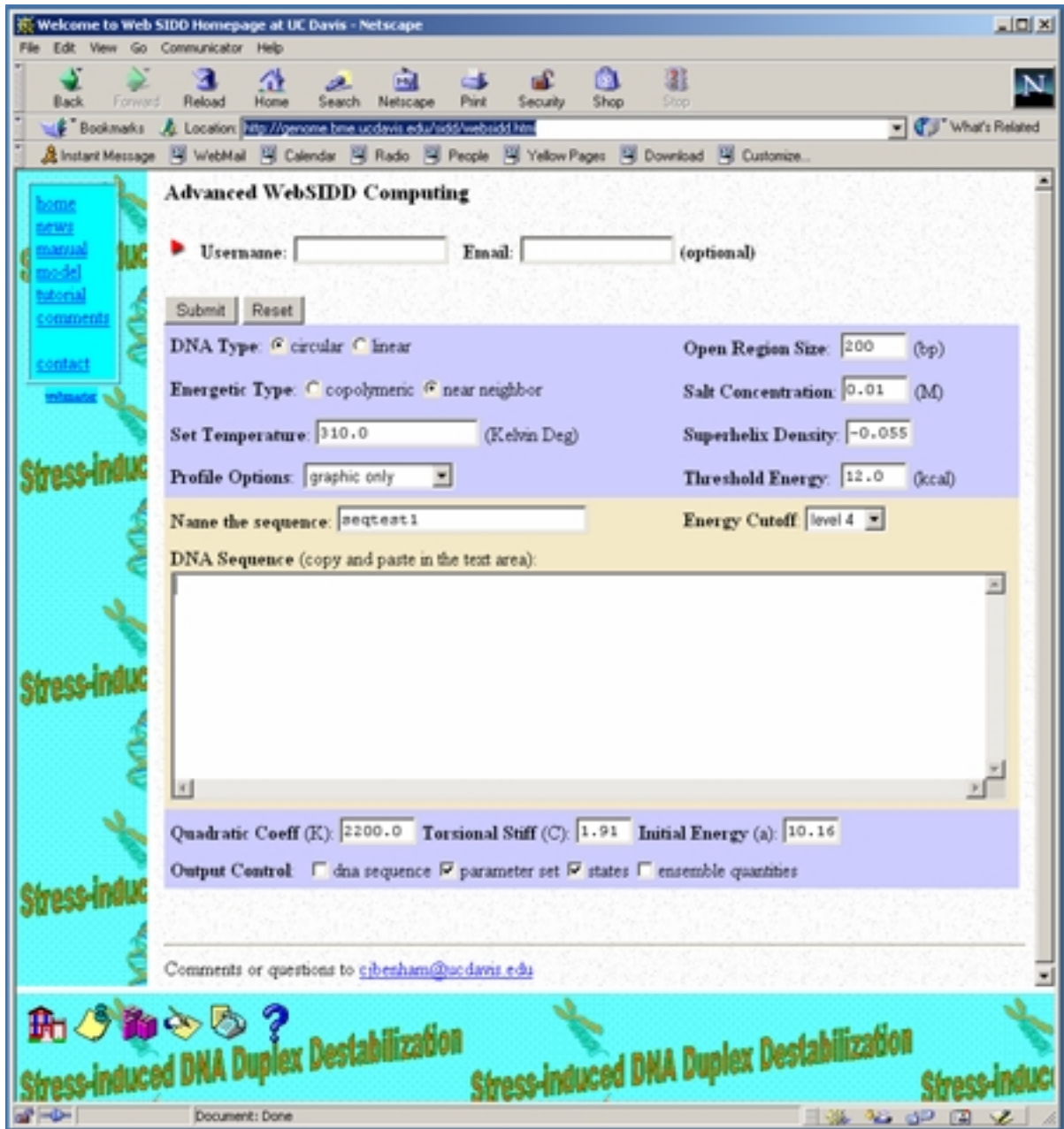


Figure 2

