Journal of Applied Mathematics and Computation, 2020, 4(3), 43-56
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

A Dynamic Model of Superhelical DNA Denaturation

Cheryl L. Sershen

Department of Computational and Applied Mathematics, George R. Brown School of Engineering, Rice University, 6100 Main Street, Houston TX 77251, USA.

How to cite this paper: Cheryl L. Sershen. (2020) A Dynamic Model of Superhelical DNA Denaturation. *Journal of Applied Mathematics and Computation*, 4(3), 43-56. DOI: 10.26855/jamc.2020.09.001

Received: June 7, 2020
Accepted: June 29, 2020
Published: July 7, 2020

***Corresponding author:** Cheryl L. Sershen, Department of Computational and Applied Mathematics, George R. Brown School of Engineering, Rice University, 6100 Main Street, Houston TX 77251, USA.
Email: csershen82@gmail.com

Abstract

Existing models of structural transitions in DNA only analyze their equilibrium properties. But the biological environment within a cell is inconstant dynamic flux. The DNA molecule is affected by numerous biological processes such as protein binding, transcription, replication, recombination, and repair. Equilibrium-based models only portray molecular properties in the thermodynamic limit and do not reflect the near-term dynamic effects of such events. Developing accurate non-equilibrium dynamic models is essential to understand these effects.
Here we present a dynamic statistical mechanical model of the response of a DNA molecule to superhelical stresses in an evolving biological environment. The master equation developed here allows us to treat secondary structural transitions, which have not been considered in previous non-equilibrium statistical mechanical studies of DNA. We focus specifically on superhelical denaturation, extending to a non-equilibrium context the approach taken by the equilibrium stress-induced duplex destabilization (SIDD) model. Our model is implemented as a time-dependent simulation using Glauber dynamics. The measures calculated include time-series distributions, such as the time-dependent energies and probabilities of opening for each base pair in the DNA molecule. Our approach enables the development of more nuanced models of in vivo DNA regulatory mechanisms and other normal and pathological processes.

Keywords

DNA dynamics, DNA structural transitions, DNA supercoiling

1. Motivation

In vivo the strands of the DNA duplex must locally separate in order to initiate transcription and replication. Certain regulatory mechanisms involve single strand-specific DNA binding proteins, which require unpaired regions to act. Even though many strand separation events are regulated by interactions with other molecules, their rates can be modulated by imposed negative superhelicity. This happens because changes of superhelicity can affect the local stability of the duplex, altering the energy cost of strand separation at susceptible sites. Sufficient negative superhelicity can drive strand opening at the most susceptible sites, and substantially decrease the energy needed for opening at others. These effects can greatly alter the rates of processes in which strand separation is a rate-limiting step.

Negative superhelicity can drive local transitions to other DNA conformations that are less twisted in the right-handed sense than is the B-form. This occurs because these transitions localize some of the superhelicity as the change of twist consequent on transition. This allows the other regions of the DNA domain to fractionally relax by a corresponding amount. If the superhelical free energy returned by this relaxation exceeds the cost of the transition, then transition will be favored at equilibrium. Superhelically driven transitions to several types of alternate DNA have been documented in vitro,

and sometimes also in vivo. These include transitions to strand separated, Z-form, cruciform and H-form conformations.

This process has been suggested to be involved in vivo processes because DNA superhelicity is known to be modulated within cells. Prokaryotes have negatively supercoiling gyrase and relaxing topoisomerase enzymes, the dynamic balance of whose activities determines a basal level of negative superhelicity. Because the gyrases are ATPases and the topoisomerases are not, the level of superhelicity produced in this way varies with the energy charge of the cell, and hence in E. coli between stationary phase and growth phase. There is considerable evidence that this change of superhelicity when transitioning between phases contributes to altering the global pattern of gene expression. In addition, transcription drives local superhelicity, with the polymerase complex leaving a wake of negative supercoils behind and producing a bow wave of positive supercoils ahead. Although most eukaryotes do not have negatively supercoiling gyrases, they do experience transcriptional superhelicity. This has been shown to travel kilobase distances and to persist long enough to drive structural transitions in vivo. Here also the extent of supercoiling affects gene regulation and other biological processes, including DNA repair [1].

Of the various alternate structures DNA can assume, only the strand separated form is presently known to be required in vivo. Both transcription and replication, the two central jobs of DNA, require helical opening, which forms the strand separated [2], [3], [4]. So changes of DNA superhelicity can alter the propensity for these open regions to form. It acts at susceptible sites to destabilize the duplex, and if strong enough can even cause strand separation [5]. Binding of specific proteins induces deformations in the DNA molecule that cause the transcription start site to denature, which alters the energetics of opening of the surrounding sites [6]. Abasic sites in DNA (a common form of endogenous damage) result in local instability of the duplex and greater flexibility [7]. Infectious agents (such as viruses) use enzymes that bind to host DNA, exerting forces that subsequently unwind it [8], [9].

The stress-induced duplex destabilization (SIDD) model is a particularly successful method to analyze the strand separation transition in superhelical DNA [6], [10]. That model uses a statistical mechanical Ising-like framework to predict the equilibrium denaturation (free) energy and probability of denaturation for each base pair in a DNA sequence. When applied to genomic DNA sequences it has shown that sites which are destabilized by superhelical stresses do not occur at random, but rather coincide with specific regulatory regions, such as transcription start sites, terminators and eukaryotic scaffold attachment regions [6], [11], [12]. The local level of destabilization in a region, given by its denaturation free energy profile, is strongly correlated with the frequency of double-strand breaks occurring there. In fact, it is a better predictor than is the GC content of the surrounding regions [13]. This result provides a strategy for predicting the recombination rates of regions based on their calculated stabilities within their genomic context.

Despite the success of this equilibrium superhelical stress model, it cannot fully describe the processes that occur in vivo, since they are not at equilibrium. Non-equilibrium methods are required to model the influence of external forces such as enzyme-DNA interactions, protein binding, dynamically imposed transcriptional superhelicity, and DNA-chemical interactions, and to describe the behavior of DNA domains after their equilibrium states are disrupted. To further understand the role that stress-induced duplex destabilization plays in biological processes and events, a computational model is needed that describes the dynamic behavior of a DNA molecule when its equilibrium is perturbed. A model of the transient openings and re-annealings that can occur is necessary to better depict both the regulatory functions of DNA and the responses of the molecule to the dynamically changing stresses that it experiences in vivo.

2. The SIDD Equilibrium Model

An equilibrium statistical mechanical model of a system is built by identifying the states accessible to the system and their associated energies. In the SIDD model one specifies the sequence of the DNA and its superhelicity, the number α of superhelical turns that are imposed on it. If the molecule contains N base pairs, it has 2^N states of strand separation. We identify the state of secondary structure of the system by specifying the values of N binary variables σ_j, j = 1,...,N, with

$$\sigma_j = \begin{cases} 0, & \text{if base pair } j \text{ is closed;} \\ 1, & \text{if base pair } j \text{ is open.} \end{cases} \tag{1}$$

These parameters correspond to the spins in an Ising model interpretation (with the caveat that they can only assume values of 0 and 1, not -1 and 1).

The transition of each base pair from the B-form to the unstressed separated state unwinds A_B = 1⁄10.4 turns of the double helix. This changes the effective superhelicity from α to $\alpha + nA_B$, where $n = \sum_i \sigma_i$ is the number of separated base pairs in that state. In addition, because single strands of DNA are relatively flexible, if there remains any torque on the ends of a denatured bubble, each pair of bases in it can helically wind by τ_i radians. So the residual superhelicity α_r, which is the portion of the imposed superhelicity that remains in this state to stress the molecule, is

$$\alpha_r = \alpha + nA - \sum_{i=1}^{N} \frac{\sigma_i \tau_i}{2\pi}. \tag{2}$$

The free energies associated with superhelicity and with interstrand twisting are both known to be quadratic. The free energy of transition has two components. Opening a new region of strand separation requires a substantial nucleation energy a = 10.2 kcal/mol. This may be regarded as the cost of creating two junctions between the open region and its B-form flanks. In addition, each base pair has an opening energy b_i, i = 1,...,N, which depends on the base identity and varies with environmental conditions, and slightly also with the identity of the flanking neighbor bases[1].

The Hamiltonian of the SIDD model derives from these considerations [6]. If σ denotes a state of this closed circular DNA, and σ_i the states of each base pair, the Hamiltonian is:

$$H(\sigma) = C(\sigma) + \sum_{j=1}^{N}[(a+b_j)\sigma_j - a\sigma_j\sigma_{j+1}]. \tag{3}$$

(We use a circular boundary condition, so that j = N + 1 corresponds to j = 1). The modification of this approach to treat linear DNA segments has been discussed elsewhere. The sum in this equation evaluates the strand separation transition energies associated with the state, while the first term is the energy associated with the residual superhelicity and the torsional interwinding:

$$C(\sigma) = \underbrace{\sum_{j=1}^{N}\frac{c\sigma_j\tau_j^2}{2}}_{\text{intertwining of unpaired strands}} + \underbrace{\frac{K}{2}(\alpha + \frac{n}{A} - \sum_{j=1}^{N}\frac{\sigma_j\tau_j}{2\pi})^2}_{\text{residual superhelicity}}. \tag{4}$$

Although this term corresponds to the field strength of the Ising model, here it is not constant but varies with the configuration of the bases.

If the set of all possible states is denoted S = σ, then the partition function of this system is:

$$Z = \sum_{\sigma\in S} Q(\sigma)\exp^{-\beta\sum_{j=1}^{N}(a+b_j)\sigma_j - a\sigma_j\sigma_{j+1}}, \tag{5}$$

where

$$Q(\sigma) = \prod_{j=1}^{N}\int_{-\infty}^{\infty} d\tau_j \exp\left(-\beta C(\sigma)\right). \tag{6}$$

The equilibrium probability $P_{eq}(\sigma)$ of any individual state of the system is

$$P_{eq}(\sigma) = \frac{\exp^{-\beta H(\sigma)}}{Z} \tag{7}$$

From this one may evaluate equilibrium ensemble average values of any quantities of interest.

Most importantly, the equilibrium probability of opening at base pair j is given by

$$P_{eq}(j) = \bar{\sigma_j} = \sum_{\sigma:\sigma_j=1} P_{eq}(\sigma) = \sum_{\sigma}\sigma_j P_{eq}(\sigma), \tag{8}$$

where the summation is over all states with $\sigma_j = 1$. In the Ising model interpretation the probability of opening at base pair j is simply σ_j, the average spin of base pair j. This is found by summing the equilibrium probabilities of all states in which base pair j is open. The opening profile of the system is the graph of $P_{eq}(j) = \sigma_j$, j = 1,...,N over the entire sequence.

The destabilization energy of base pair j is defined to be difference between the weighted average of the energy of opening at base pair j, minus the ensemble average free energy of the system G:

$$G(j) = \frac{\sum_{\sigma:\sigma_j=1} H(\sigma)\exp^{-\beta H(\sigma)}}{\sum_{\sigma:\sigma_j=1}\exp^{-\beta H(\sigma)}} - \bar{G} \tag{9}$$

The values of all the energy and structural parameters have been determined from experimental data. Here $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant and T is the temperature in degrees Kelvin. Also c $\cong$ 9.3 ×10^{-21} kcal/mol/rad^2 is the torsional stiffness of interstrand winding of the two strands in a separated state. The nucleation free energy is a $\cong$ 10.16 kcal/mole. The coefficient of the quadratic free energy associated to the residual linking difference is K $\cong$ 2220RT/N [14]. The helical twist of B-form DNA is A = 1/10.4 turn/bp [15]. The opening energy b_j of base pair j can be computed

[1]Both near-neighbor and copolymer energetics have been explored in the SIDD model.

using nearest neighbor energetics or by copolymer methods. The latter case yields $b_{AT} \cong 0.26$ kcal/mole and $b_{GC} \cong 1.31$ kcal/mole under the conditions modeled here [14].

The standard SIDD model is implemented in an approximate algorithm, in which equilibrium values are estimated by averaging over those states whose energies fall within a specified threshold above of that of the minimum free energy state. Experience has shown that this approach provides a satisfactory balance between computational efficiency and accuracy.

3. The Kinetic Model

The kinetic model developed here simulates the time evolution of a collection of M identical molecules, each with the same base sequence and level of imposed superhelicity. Each base pair can occur in either of two configurations, B-form or strand separated. So if the number of base pairs in the molecule is denoted by N, each molecule has 2^N possible conformational states. A state σ is determined by specifying the values of N binary variables σ = (σ1,...,σN), with σi = 1 if base pair i is open and $\sigma_i = 0$ otherwise. By imposing the periodic boundary condition $\sigma_1 = \sigma_{N+1}$ one may regard the molecules as being circular, as is sometimes done below.

Here we develop a discrete time model and follow the population of molecules as it approaches its equilibrium distribution. The binary variables describing the conformations of each of the base pairs then become functions of time, $\sigma_i(t)$. The M molecules are started either in identical initial states or in a variety of randomly selected states. The time evolution of the system is modeled by specifying how the molecular conformations change between successive time steps. This analysis uses the techniques of Glauber Dynamics, described in general next.

3.1. Glauber Dynamics

Glauber dynamics have been used to model the approach to equilibrium of a one-dimensional Ising Model [16], [17], [18]. There the state $\sigma(t) = [\sigma_1(t),...,\sigma_N(t)]$ describes the orientations of the spins along the chain of length N. The spin state at each position is treated as a stochastic function of time, with a possibly position-specific transition probability function $w_i(\sigma_j(t))$. This function also may depend on the current conformational state of the molecule. In the dynamic SIDD model $w_i(\sigma_j(t))$ is the instantaneous probability that the jth base pair of configurationsi changes from one state to the other within a time step, while all other sites remain fixed.

Because multiple simulations are followed, one determines the frequency distribution of them among the possible states at each time step. Suppose this distribution at time t is $P(\sigma_1(t),...,\sigma_N(t)) = P(\sigma(t))$. In a discrete Glauber dynamics model the components of the distribution function at time t obey the (discrete time) differential equation:

$$P(\sigma_1,...,\sigma_N,t+1) = P(\sigma_1,...,\sigma_N,t) \underbrace{-[\sum_j w_i(\sigma_j)]P(\sigma_1,...,\sigma_N,t)}_{\text{annihilation of } \sigma=(\sigma_1,...,\sigma_N)} + \underbrace{\sum_i w_i(\sigma'_j)P(\sigma_1,...,\sigma'_j,...,\sigma_N,t)}_{\text{nucleation of } \sigma \text{ from } \sigma'=(\sigma_1,...,\sigma'_j,...,\sigma_N)} = WP \tag{10}$$

where W is the 2^N X 2^N self-adjoint matrix of transition probabilities and P is the 2^N vector of time dependent probability functions. The result follows since the matrix W is the right-stochastic matrix of transition probabilities with $\sum_j w_i(\sigma_j) = 1$. Here $\sigma_j'$and σ_j denote the two values of the binary variable describing the spin at position j (or, below, the conformation of the j-th base pair). The expression in Eq. (10) is referred to as the master equation.

This simple Ising model is transformed into a kinetic model by choosing the transition functions $w_i(\sigma_j)$ to give probability dynamics that converge to the equilibrium distribution:

$$lim_{t\to\infty} P(\sigma_1,...,\sigma_N,t) = P_{eq}(\sigma_1,...,\sigma_N) \tag{11}$$

In equilibrium, $WP_{eq}(\square_1,...,\square_N) = P_{eq}(\square_1,...,\square_N)$. This requirement will be satisfied if the transition functions are chosen so the system satisfies the conditions of detailed balance and ergodicity. Detailed balance condition means that,

$$w_i(\sigma_j)P_{eq}(\sigma_1,...,\sigma_N) = w_j(\sigma'_i)P_{eq}(\sigma_1,...,\sigma'_j,...,\sigma_N). \tag{12}$$

The ergodicity condition requires the state space of the system to be (pathwise) connected, and the states are visited aperiodically. For the state space to be connected, it must be possible to reach any configuration from any other configuration with finite, non-zero probability. Aperiodicity means the state space does not have any subsets that are visited cyclically. If these conditions are satisfied, then the matrix W of transition probabilities $w_i(\sigma_j)$ has real eigenvalues, and the

eigenvalue $\lambda_0 = 1$ corresponds to the equilibrium distribution. Relaxation of physical quantities of interest may be expressed in terms of the eigenvalues and eigenvectors of W [17].

One case where the Master equation is easily solved is the ferromagnetic simple Ising model with near-neighbor cooperativity and zero external field, where $\sigma_j(t) = \pm 1$ [17]. In practice, it often is a formidable challenge to solve the Master equation when the external field is non-zero. In these cases it is much more difficult to ensure that the detailed balance, aperiodicity and connectedness conditions are met. Some results in this area have been developed, most notably in [19].

The approach developed here to model SIDD dynamics in superhelical DNA is novel in that it solves the Master equation in the presence of not simply near-neighbor, but of global coupling among the base conformations $\sigma_j(t)$ imposed by the quadratic energy of residual superhelicity, as described below.

3.2. A Dynamic SIDD Model

In our dynamic version of the SIDD model, the spins correspond to the conformations of the base pairs:

$$\sigma_j(t) = \begin{cases} 0, & \text{if base pair } j \text{ is closed;} \\ 1, & \text{if base pair } j \text{ is open.} \end{cases} \tag{13}$$

The Hamiltonian governing the SIDD transition contains an 'external field' term arising from the nucleation and transition energies for strand separation. It also contains a stress energy term, $C(\sigma(t))$, that includes the free energy of the residual superhelicity and the torsional deformation energy of the strand separated regions. Following the example of Glauber dynamics, we introduce the time-dependent Hamiltonian that governs each molecule in the population:

$$H(\sigma(t)) = \underbrace{C(\sigma(t))}_{\text{stress energy}} + \underbrace{\sum_{j=1}^{N}(b')\sigma_j(t)}_{\text{nucleation + binding energy}} - \underbrace{a\sigma_j(t)\sigma_{j+1}(t)}_{\text{near-neighbor (stacking) energy}} \tag{14}$$

where $b' = (a + b_j)$, and

$$C(\sigma(t)) = \underbrace{\sum_{j=1}^{N}\frac{c\sigma_j(t)\tau_j(t)^2}{2}}_{\text{torsional deformation energy}} + \underbrace{\frac{K}{2}(\alpha + \frac{n(t)}{A} - \sum_{j=1}^{N}\frac{\sigma_j(t)\tau_j(t)}{2\pi})^2}_{\text{energy of residual superhelicity}}, n(t) = \sum_{j=1}^{N}\sigma_j(t) \tag{15}$$

The last term in Eq. (14) implicitly assumes the molecule is circular, so $\sigma_{N+1} = \sigma_1$. Also, the total number n(t) of base pairs in the open state in a molecule at time t is

$$n(t) = \sum_{j=1}^{N}\sigma_j(t). \tag{16}$$

The presence of the stress energy $C(n(t),\sigma(t))$, Eq. (15), introduces several complications into the system. First, the term in Eq. 15 giving the energy of residual superhelicity contains twist parameters $\tau_j(t)$ that must be evaluated. Here we assume that all open base pairs have the same twist $\tau_j = \tau$, and select the value that minimizes the Hamiltonian. (Base-specific methods to handle this issue also have been developed [6], but they have been shown to give essentially equivalent results as the simpler strategy presented here.) Since only the superhelical stress term $C(\sigma(t))$ contains components that involve τ, this is equivalent to calculating:

$$\frac{\partial C(\sigma(t))}{\partial \tau} = 0 \tag{17}$$

performing this calculation and solving for τ. Taking the derivative with respect to τ and setting it equal to 0, finds the minimal value of τ to depend on the total number n of open bases as:

$$\tau(n(t)) = \frac{2\pi K(A\alpha + n)}{A(4\pi^2 C + Kn)}. \tag{18}$$

Further, the last term in Eq. 15, which describes the free energy of residual superhelicity, includes terms in which each spin is coupled to every other spin. This global coupling renders the SIDD phenomenon both more complicated to analyze and potentially much richer in the repertoire of behaviors it can exhibit. As $\sigma_j^2 = \sigma_j$, expanding this term yields:

$$\frac{K}{2}[\alpha^2 + \frac{1}{4A^2\pi^2}\sum^{N}\sigma_j(8A\alpha\pi^2 + 4\pi^2) - (\alpha 4A^2\pi + 4\pi A)\tau + A^2\tau^2 + \underbrace{\frac{1}{2A^2\pi^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\sigma_i\sigma_j(4\pi^2 + A^2\tau^2) - \frac{1}{A\pi}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\sigma_i\sigma_j\tau}_{\text{terms in which each spin is coupled to every other}}]. \tag{19}$$

So this SIDD model contains both a standard 'effective external field' term, and a term in which the conformation of any base pair is coupled to the conformations of all other base pairs in the molecule. This latter term distinguishes the SIDD model from the standard Ising model in a magnetic field, which only includes near neighbor influences.

In order to implement the dynamic SIDD model using Glauber dynamics, we need to determine the transition probability functions $w_i(\sigma_j)$. This may be done by considering a single time step. There are only two states to consider in each time step, which we call σand σ′and which differ by a single "spin", say σ_j, the spin at position j. Now, for our model to converge to equilibrium it must satisfy the detailed balance condition described in Eq. (12). Expressing the equilibrium probabilities of the system in terms of Boltzmann factors gives:

$$\frac{w_i(\sigma_j)}{w_j(\sigma_i')} = \frac{P_{eq}(\sigma_1,...,\sigma_j',...,\sigma_N)}{P_{eq}(\sigma_1,...,\sigma_N)} = \frac{\frac{\exp^{-\beta H(\sigma')}}{Z}}{\frac{\exp^{-\beta H(\sigma)}}{Z}} = \frac{\exp^{-\beta H(\sigma')}}{\exp^{-\beta H(\sigma)}}$$

$$= \exp^{-\beta(H(\sigma')-H(\sigma))} = \exp^{-\beta\Delta H}, \tag{20}$$

where ΔH = H(σ′) -H(σ). Because we need only consider these two states, we may form a reduced partition function from their energies, and calculate their transition probabilities as follows:

$$w_i(\sigma_j) = \frac{\exp^{-\beta H(\sigma')}}{\exp^{-\beta H(\sigma)} + \exp^{-\beta H(\sigma')}} = \frac{1}{1+\exp^{\beta\Delta H}},$$

and

$$w_j(\sigma_i') = \frac{\exp^{\beta\Delta H}}{1+\exp^{\beta\Delta H}}, \tag{21}$$

This shows that Eq. 18 is satisfied, so detailed balance holds.

It is important to note that, although only the conformation at position j is allowed to change in this step, the values of the Hamiltonians H(σ′) and H(σ) depend on the configurations of all other base pairs. So the transition probabilities $w_i(\sigma_j)$ are not fixed for each j, but instead also depend on the behaviors of the other base pairs, as they must in a globally coupled context.

To show that ergodicity also holds, one must demonstrate that the connectedness and aperiodicity conditions are satisfied. Connectedness follows from the observation that the probabilities of a given base pair flipping or remaining the same are both always positive. As any configuration can be reached from any other by a finite number of flips (for given strand length of finite N), each of which has positive probability, this proves connectedness. Aperiodicity is satisfied because every state has a positive probability of being reached in one step from itself. That is, the system has positive probability of remaining in the same state in the next timestep, so no state can be in a periodic subset [17].

3.3. Algorithmic Implementation of the Dynamic SIDD Model

The dynamic SIDD model is implemented by using a Monte Carlo strategy to analyze a set of M identical molecules as they pass through successive time steps. Their initial states may be chosen in whatever way suits the problem being analyzed. At each time step each molecule is considered individually as follows. First, the base pair j whose transition is to be considered is selected at random. The transition probability $w_i(\sigma_j)$ of that base pair is determined from the Hamiltonians of the initial and the ending configurations, as described above. Then a second random number $0 \leq r \leq 1$ is chosen, and the transition is performed if $r < w_i(\sigma_j)$. Otherwise the molecule remains in its initial state. These steps are carried out for each molecule in the set before moving to the next time step.

Empirical probabilities may be calculated for any quantities of interest from the simulation results. The probability of any state σ at time step t is estimated to be:

$$P(\sigma,t) = \frac{\text{number of times } \sigma \text{ is reached at time step } t}{\text{total number of molecules simulated}} \tag{22}$$

The probability that base pair j is open at time step t is

$$P(\sigma_j = 1,t) = \sum_{\sigma}(P(\sigma,t)\sigma_j). \tag{23}$$

Other quantities that can be calculated include the ensemble average energy G(t) of the system at each time step t, number of runs of open base pairs and the average length of a run, at time step t. The destabilization energy of base pair j is related to the average energy of all states in which that base pair is open:

$$G(j,t) = \frac{\sum_{\sigma} H(\sigma(t))\sigma_j}{\text{number of states with base pair } j \text{ open at time step } t} - \bar{G}(t). \tag{24}$$

If there are no occupied states in which base pair j is open, one infers that their Hamiltonians are all so large that they are not sampled in the limited population being considered. In that case one may say that the destabilization free energy lies beyond a (large) threshold value. Relaxation times and rates of approach to equilibrium be estimated from this data by calculating the curves of best fit to the simulated data and comparing with experimental results.

3.4. Parameter values

In this implementation, the transition energetics are assumed to be copolymeric. This approach assigns transition energy b_{AT} = 0.26 kcal/mol to every A·T base pair, and b_{GC} = 1.31 kcal/mol to every G·C base pair. These are the values that have been experimentally measured at a temperature of 310K, as cited in Section II. The linking difference is calculated from the supercoiling density as: $\alpha = \frac{\sigma N}{A}$, where A=10.4 base pairs per turn is the helical repeat of unstressed B-form DNA and N is the DNA sequence length. The coefficient K in equation 20 was experimentally found to be 2220 RT/N under moderate salt conditions, where R is the gas constant and T is temperature. The torsional stiffness c of Eq. (3) is 1.91 kcal/mole/rad^2. The nucleation energy a is the energy needed to nucleate a run of open base pairs. Its value is dependent on temperature and ionic strength. In the Glauber program, we set a = 10.16 kcal/mole. The reasons for these choices have been presented elsewhere [14].

3.5. Algorithm Parallelization

The implementation was developed on the Shiraz high performance cluster at the U.C. Davis Genome Center with 111 available nodes. The simulation and statistics components were written in C++ in two separate parts, ordered and called by a Python script that uses the qsub job scheduler to submit and coordinate the simulation routines to run in parallel, and combines the interim output files. An additional C++ statistical module calculates the aforementioned quantities of interest. The workflow of the parallelized version of this code is summarized in Figure 1.

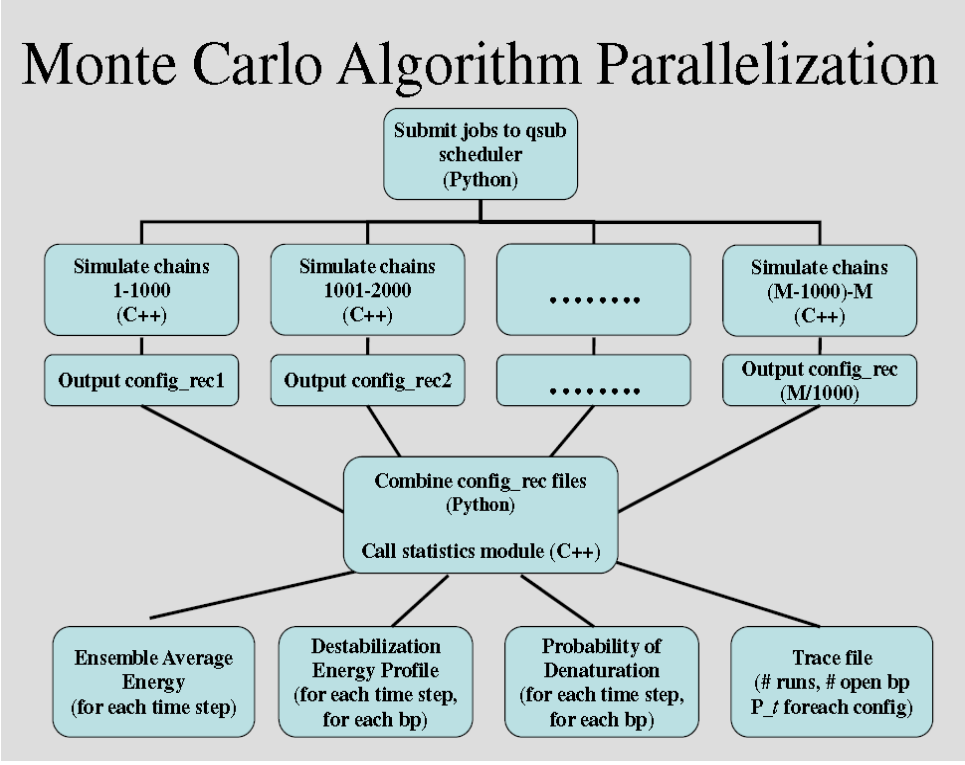


Figure 1. Running Dynamic SIDD in Parallel.

4. Results

4.1. Comparison with the Equilibrium SIDD Model

We show the near equilibrium profile for the dynamic model using the Glauber algorithm versus the equilibrium SIDD profile for the sequence listed in Figure 2. The two model profiles are in agreement when the ensemble average energy of the dynamic model is about 10 kcal/mole. This arises from the fact that approximate method of the equilibrium SIDD algorithm uses a threshold of 10 kcal/mole when deciding what states to consider in calculating the equilibrium. The two profiles are in close agreement, profiles with ensemble average energy equal to 8.2291 Kcal/mole and 9.139 Kcal/mole failed to reject the null in a KS test with $p < 0.0250$. To run this test and create the figure, 10,000 replications of the strand were run with random initial probability of opening equal to 0.35 for each base pair in the chain with superhelix density equal to -0.055.

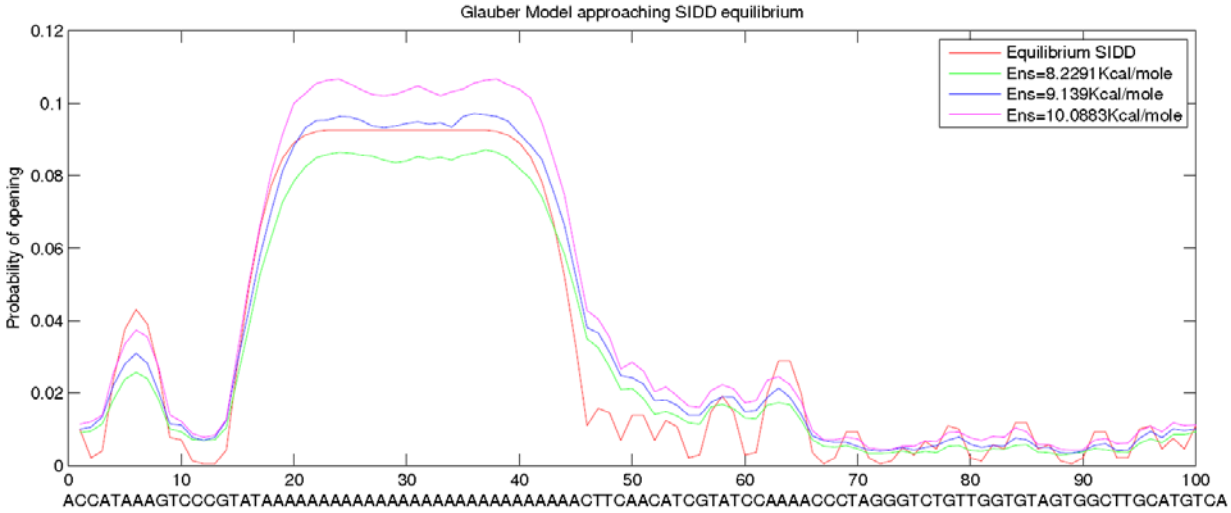


Figure 2. Comparison of the Glauber probability of opening profiles with the equilibrium SIDD probability of opening profile.

4.2. Varying Molecular Length

We show in Figure 3(a) that DNA strands of length 100 base pairs converge to the state in which all base pairs are closed, since the number of open base pairs that minimizes the Hamiltonian for this strand length is equal to zero. But the system is actually a probability distribution over all the states, with states with open base pairs having very low probability for the case N=100. Since there are $2N = 2^{100} = 1.26765E + 30$ configurations for the 100 base pair chain, it is not practical to compute the complete partition function in order to calculate the equilibrium probability of the closed state.

Instead, we select a homopolymer strand and perturb it slightly. Given the strand length, the optimal number of open base pairs is found by solving the minimization problem:

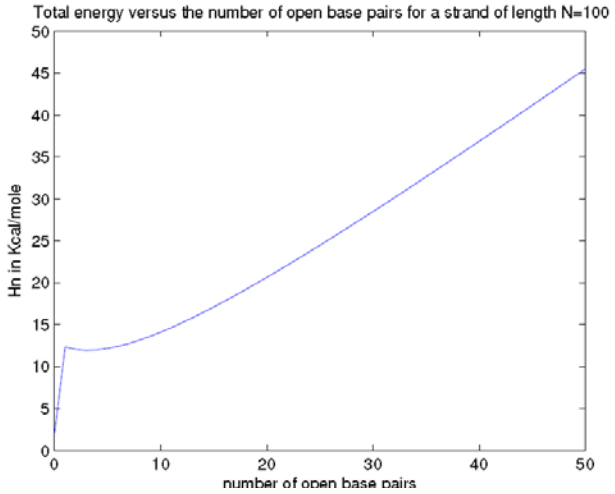
$$\frac{\partial H}{\partial n} = 0 \tag{25}$$

since the equilibrium probability equation assigns a penalty as the energy of the Hamiltonian increases, so that:

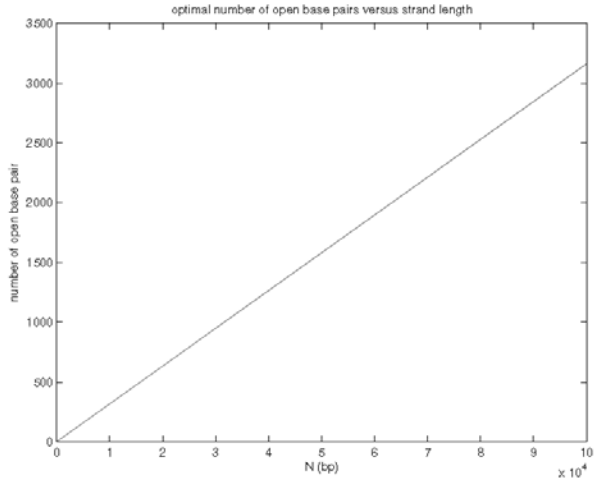
$$n = \frac{-(\frac{16\pi^4 C^2 K}{A^2} + 8\pi^2 CKb) + \sqrt{((\frac{16\pi^4 C^2 K}{A^2} + 8\pi^2 CKb)^2 - 4 * (K^2 b + \frac{2\pi^2 CK^2}{A^2}) * (16\pi^4 C^2 (\frac{K\alpha}{A} + b) - 2\pi^2 CK^2 \alpha^2)}}{2 * (K^2 b + \frac{2\pi^2 CK^2}{A^2})} \tag{26}$$

In Figure 3, we first plotted the energy surface for the strand with 100 open base pairs (n) and see that the number of open base pairs that minimizes the homopolymeric energy function for a single run is n=0 open base pairs. This shows that there exists an energy deficit at n = 0 which must be overcome before equation 26 becomes the governing equation (see figure 3c, where we plot the energy of n=0 versus n determined by equation 26). For strands of length less than N = 1,254, the energy of the all closed state minimizes the Hamiltonian. Figure 3d shows the entire energy surface, as determined by equation 26. For strands of length N ≥1,254, equation 26 pertains.

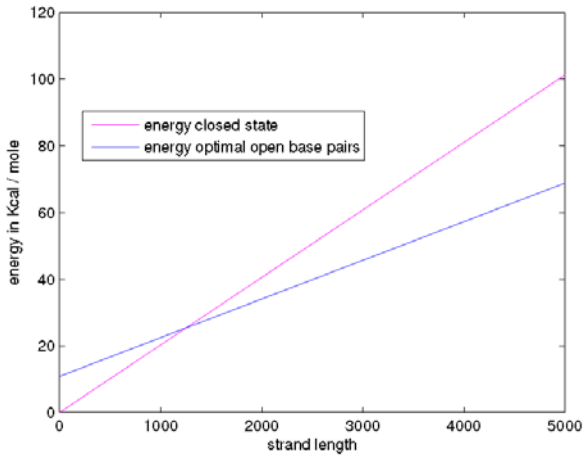
Table 1 shows the effect of varying molecular length on the number of time steps to convergence. Each run used supercoiling density of -0.055, 1,000 molecules and used 4 nodes. The number of time steps to convergence does not increase significantly until molecular length exceeds 1,000 base pairs. The effect is modest linear increase.



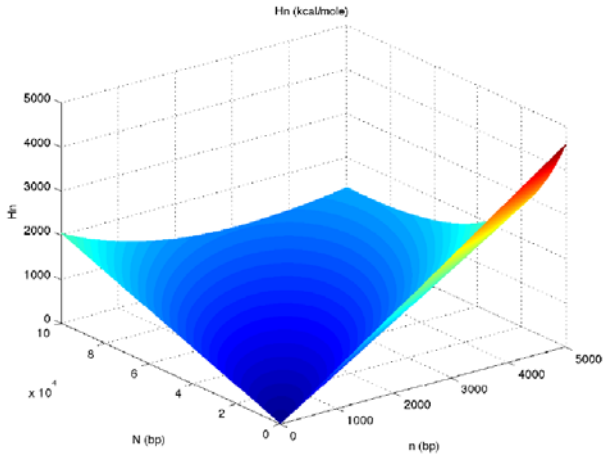
(a) a



(b) b



(c) c



(d) d

Figure 3. All calculations are for a homopolymer strand with one run of open base pairs. Figure a: the energy surface for the N = 100 base pairs vs n open base pairs. n = 0 open base pairs is still the lowest energy configuration. Figure b: the optimal number of open base pairs for different strand lengths, as determined by equation 26. Figure c: for strands of length 1,253 or less, the state with no open base pairs is the lowest energy state. For strands of length greater than 1,253, equation 26 gives the optimal number of open base pairs. Figure d: the energy surface for different strand lengths, for various number of runs of open base pairs.

Table 1. Effect on convergence of varying molecular length

Molecular Length (bp)	Total Time Steps to Convergence
100	34
200	32
500	31
1,000	44
2,000	186
5,000	300+

5. Applications

5.1. Transcription

We used our model to dynamically simulate the action of the transcription factor binding explored in Sheridan et al. [5].

We simulated the action of IHF binding in the 181 base pair upstream promoter region (strong SIDD site.) Figure 4 shows the time evolution of the superhelical stress profiles in the upstream promoter region. We simulated 7,500 replications of the 9,000 base pair strand at supercoiling density of -0.055 and extracted the 180 base pair IHF binding site (shown in Figure 4).

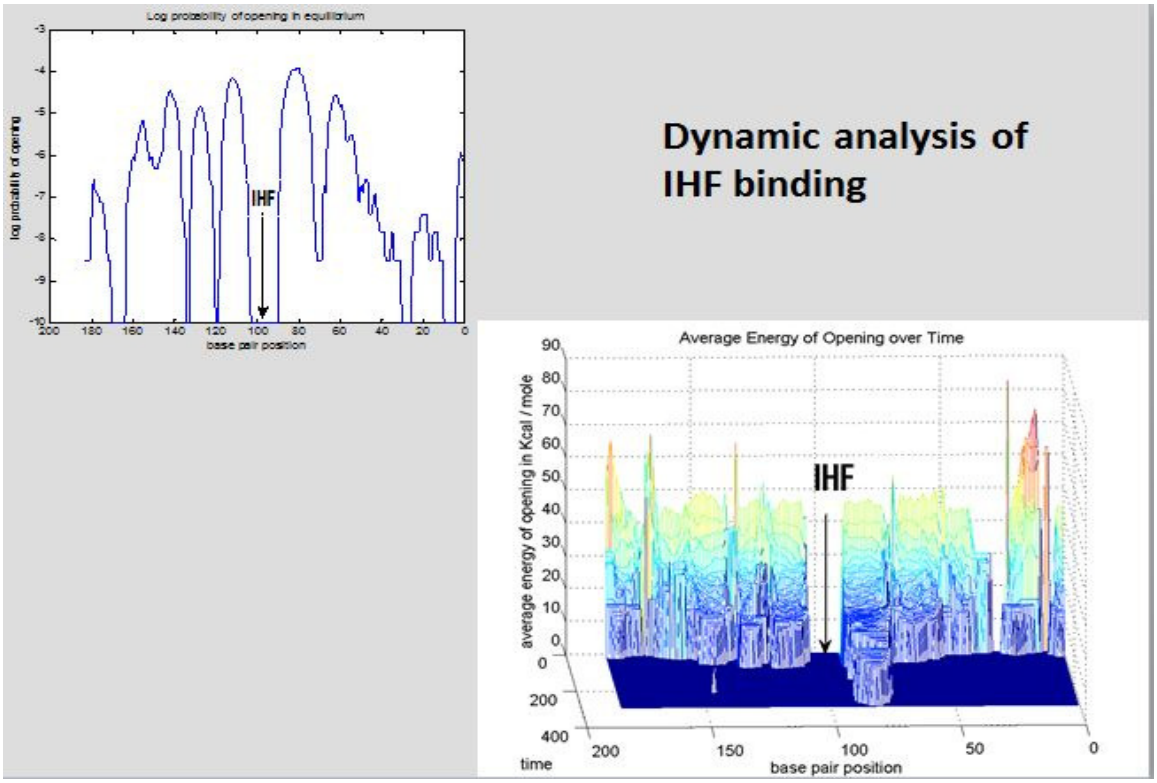


Figure 4. Probability of opening and Energy surface of the ilvPg promoter regulatory region of the ilvGMEDA operon of E. coli.

At time step 0, we apply a constant force F_i in the IHF binding region (between base pairs 80 and 100 in figure 9), by setting the chemical binding energies in this region for base pairs -80 to -100 equal to 5 kcal/mole for AT base pairs and 10 kcal/mole for GC base pairs. This simulates the forcing of these base pairs closed. At time step 5, the entire region is open, except for the IHF binding site. The region to the right is more strongly destabilizied, nearer to the transcription start site. The area near the transcription start site remains the most strongly destabilized, throughout the time evolution of the stress profiles.

The dynamics of the binding process are shown in Figure 4, which describes the changes in the energy surface of the system over time steps, over the entire stretch of the strong SIDD region, 180 base pairs upstream of the transcription start site. Binding of IHF in the $ilvP_G$ promoter region shifts the profile, alters the negative superhelical stress distribution and transiently opens up the base pairs nearer to the transcription start site, up regulating transcription. This is seen particularly by the region of strong destabilization in the -10 region of the transcription start site and the relatively high probability of opening in the -10 region. Thus the model is consistent with the findings in Sheridan et al.

5.2. Protein Binding

Another scenario that may be easily modeled under the Glauber dynamics is the case where a protein binds to a DNA structure, closing off a region and transferring stress far downstream onto competing destabilized sites. Aranda et al. [20] found an alternating $d(TA)_n$ sequence in the 3' region of the Saccharomyces cerevisiae FBP1 gene that when deleted, transferred the superhelical stress onto alternative competing nuclease hypersensitive sites. Though Aranda et al. used a plasmid structure to study the effects of deletions of the $d(TA)_n$ sequence, we extracted an 11,000 base pair region from Saccharomyces cerevisiae chromosome XII that included the FBP1 gene [and the inherent $d(TA)_n$ sequence] and approximately 5,000 base pairs downstream of the gene. Instead of deletions, we simulated protein binding at the site by binding the base pairs of the $d(TA)_n$ sequence to examine the effects on the competing SIDD sites.

Figures 5, 6, and 7 show the evolution of the destabilization energies across the chromosomal region at different time steps.

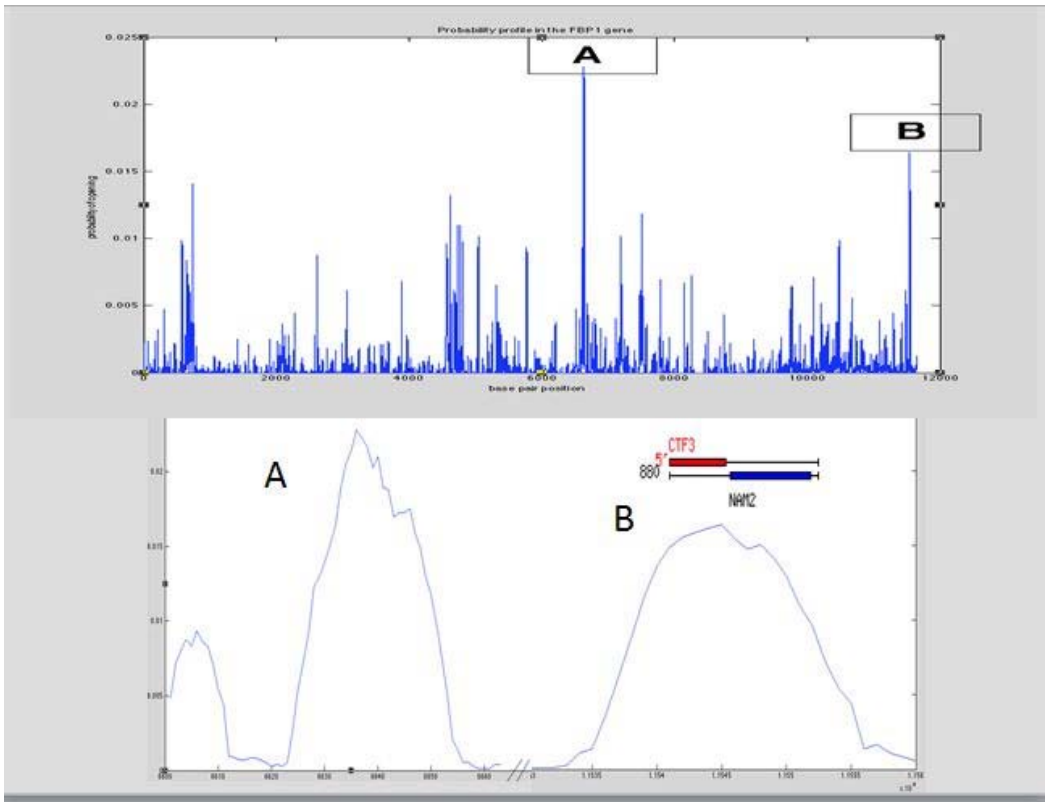


Figure 5. 11,000 base pair region including the FBP1 gene (peak A).

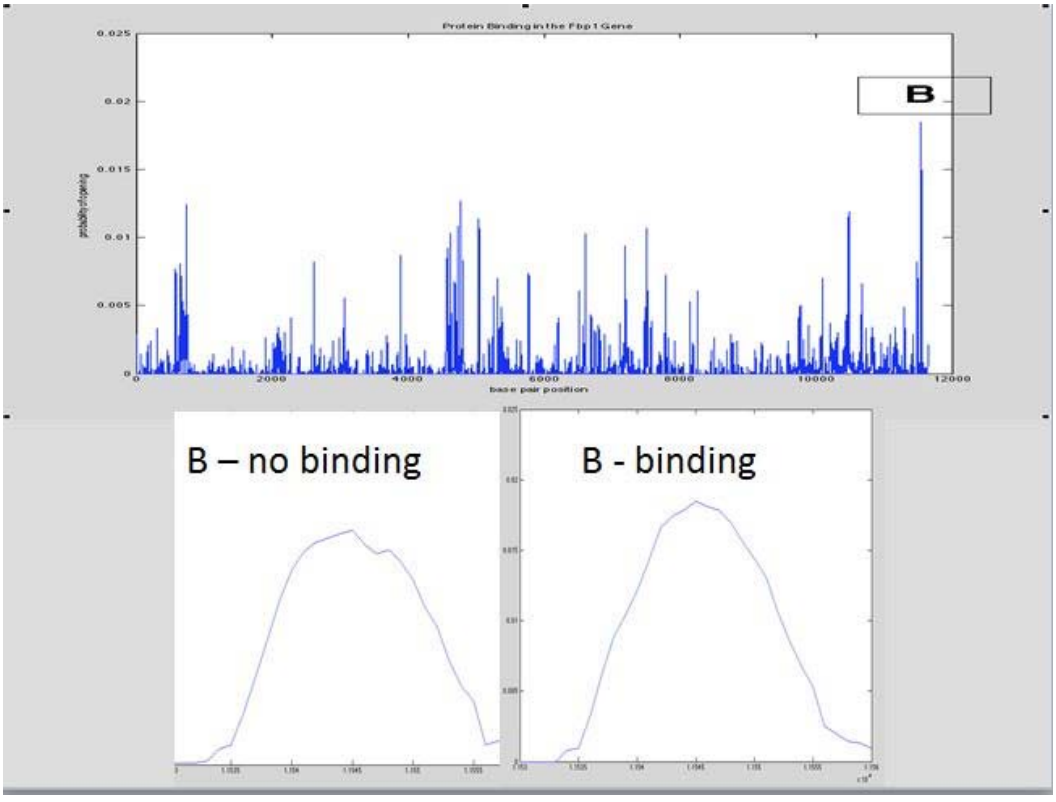


Figure 6. 11,000 base pair region with binding in the peak A region and peak B both with and out binding.

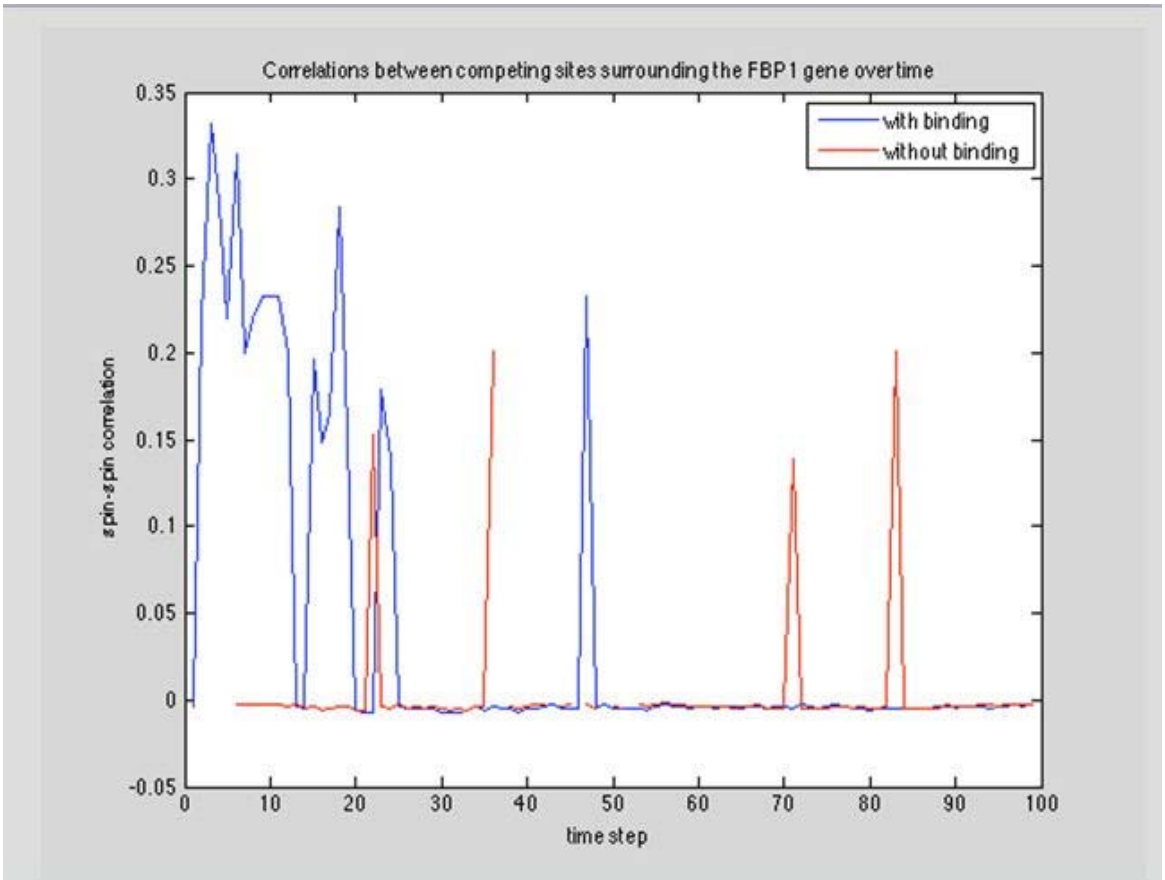


Figure 7. Spin-spin correlations between the peak A and peak B regions, both with (blue) and without binding (red).

The alternating $d(TA)_n$ sequence is in the center of figure 5 at the site labeled 'A' within the FBP1 gene. We simulated 7500 molecules at supercoiling density -0.055 and averaged the probabilities of opening. The top plots the probability of opening across the 11,000 base pair sequence. The alternating $d(TA)_n$ sequence (peak A) is the site with the highest probability of opening. A competing site arises near the end of the sequence, at a location between two ORFs (CTF3 Watson, NAM2 Crick), one on the Watson strand and one on the Crick strand, so it could coincide with a terminator site for either (or both) of the two genes.

Figure 6 shows the same region with the alternating $d(TA)_n$ sequence bound closed. Note that the competing destabilized site labeled 'B' has increased probability of opening in Figure 6. Now we bind the repeating (TA)n sequence, and loose peak A. The stress is transferred onto peak B, which now has a larger probability of opening.

Figure 7 shows the spin-spin correlations in the neighborhood of the bound region and at the center of peak B and shows that the spin-spin correlations are both significant and significantly higher when the $d(TA)_n$ sequence is bound This shows that torsional stress can be transferred many kilobases downstream onto alternating competing sites, when binding of protein or transcription factor occurs.

6. Discussion

Since the superhelical stress profiles contain much information about the internal structure of a given DNA strand, we can use the information inherent in the profiles to study the effects of dynamic biological events such as transcription and protein binding, and use the destabilization profiles to do location analysis of promoter regions, transcription start sites, protein binding sites, SMARs and viral insertion sites within the molecule.

We will use the model to study the binding of proteins to enhancer sites and the binding of factors that act as repressors of transcription. With our model it will be possible for the first time to view the energetics of the actions of such biological events. Current experiments are revealing more about the ways that proteins like zinc fingers actually bind to the DNA. The action of these external agents on the DNA molecule may be easily modeled within the framework presented here. Additionally, we may alter the Hamiltonian to reflect the energetic effects of such processes. In the case of

zinc finger binding, for example, we may add a term to the Hamiltonian to reflect the additional stresses induced by the protein winding around, twisting and bending the DNA, or relaxing it in the region where the binding takes place. We are currently collaborating other labs to model these scenarios.

The model may be used to study the binding of drugs and other molecules using the same basic principles as in the binding of transcription factor and proteins.

The model could also easily be used to study scenarios of DNA damage and repair. This problem can be addressed by allowing certain base pairs to be open in the damaged state and allowing the molecule to relax to equilibrium, or by delineating the open base pairs in the starting state (or runs of open base pairs). Here we allowed each base pair a 10%-35% chance of being open at the 0th time step and studied the convergence to equilibrium, but this is quite similar to the problem of radiation damage in DNA. It would also be useful to collaborate on damage and repair experiments in order to get a first glimpse of the timescale on which relaxation from a high energy state occurs. In this way, we may fit the model output to experimental data and determine the scale of the model's time steps.

In the present study, we have simulated only DNA strands of up to 11,000 base pairs. It would be interesting (and possible given the parallel implementation) to investigate the convergence dynamics of much larger strands. Monte Carlo sampling methods may be used to selectively sample states for very large molecules. Here we would record the information only for every nth state visited by the molecule. Also, we would like to extend our analysis of the effects of molecular length on the measures presented here.

We have developed a novel approach for analyzing the dynamics of DNA structural transitions and function. This is the first non-equilibrium statistical mechanical approach to modeling DNA and obtaining a dynamic picture of the DNA molecule using ensemble averages, destabilization energies and probability of denaturation. Previous approaches were limited to single molecule transitions and did not take into account structural transitions. Although single molecule Monte Carlo has been used extensively to model DNA dynamics, this approach has only been used once to explore structural transitions [21]. The other approaches of which we are aware used single molecule Monte Carlo simulation [22] and Brownian dynamics approaches to modeling the single DNA molecule [23]. For a review of single model DNA dynamics and protein binding see [24]. Examples using molecular dynamics simulations can be found here [25, 26]. The current study represents the first application of non-equilibrium statistical mechanics to analyze superhelical and structural transitions in DNA. Our method uses large-scale Monte Carlo simulation to provide a dynamic picture of both the ensemble and single molecule fluctuations around equilibrium. We are also able to visualize approach to equilibrium and study the mechanisms of binding and release.

The master equation developed here allows us to introduce complexities not seen in previous non-equilibrium statistical mechanical studies of the DNA molecule. The global coupling of the base pairing in the model induces physics that have not heretofore been employed in studying dynamic models of DNA.

Our model allows the user to set and dynamically change conditions in a way that is not possible with single molecule Monte Carlo simulations. This technique allows the user to examine changes induced by processes like DNA repair, transcription and protein binding. Though our results to date have focused mainly on validations and first attempts at applications, we plan to use the model to understand the biology of features of DNA such as function of binding and promoter regions, role of enhancers and repressors, and the dynamics of proteins binding such as zinc fingers.

When the model is started from an arbitrarily high energy state, the ensemble average energy of the system converges exponentially over time steps. We used the model to show that the dynamics of copolymer strands are intermediary to those of the all AT and all GC homopolymer cases. This is to be expected since it requires substantially more energy to open GC, as opposed to AT, base pairs.

For shorter DNA strands, the equilibrium probability of the state in which all base pairs are closed is very close to 1, so all base pairs in the molecule eventually close. For longer strands, the equilibrium is a distribution over states, so the destabilization and probability of denaturation profiles vary over the equilibrium distribution. As expected, the number of time steps to convergence increases linearly with molecular length, when started from a random high energy state.

7. Acknowledgments

I would like to acknowledge everyone in the Benham Laboratory at the U.C. Davis Genome Center, especially Dr. Craig J. Benham for his editorial suggestions about this manuscript. Also, Sally Madden, Juliette N. Zerick, Eva M. Strawbridge, Dr. Steve P. Mielke (formerly of the Genome Center), Dr. Tom Huang and Adam Dobrin for their invaluable suggestions for refining the dynamic model. Dr. Alfred O. Ong'iro provided valuable critiques, suggestions and support. I would like to thank the VIGRE funding committee for providing the funds that allowed us to begin this research. Mike Lewis of the Genome Center provided much guidance on using the high performance cluster at the Genome Center and was also an endless source of support for the implementation process.

References

[1] Kouzine, F. and Levens, D. (2007). Supercoil-driven DNA structures regulate genetic transactions. Front. Biosci. vol. 12, pp. 4409-4423.

[2] Blake, R. D., Bizzaro, J. W., Blake, J. D., Day, G. R., Delcourt, S. G., Knowles, J., Marx, K. A. and SantaLucia, J., Jr. (1999). Statistical Mechanical Simulation of Polymeric DNA Melting with MELTSIM. Bioinformatics, vol. 15, no. 5, pp 370-375.

[3] Yang, H., Zhuo, Y., and Wu, X. (1994). Investigation of Thermal Denaturation of DNA Molecules based on Non-equilibrium Transport Approach. J. Phys. A Math. Gen, vol. 27, pp. 6147-6156.

[4] Lubensky, D. K., Nelson, D. R. (2000). Pulling Pinned Polymers and Unzipping DNA. Phys. Rev. Lett. vol. 85, no. 7, pp. 1572-1575.

[5] Sheridan, S., Benham, C. J., and Hatfield, G. W. (1998). Activation of gene Expression by a Novel DNA Structural Transmission Mechanism That Requires Supercoiling-induced DNA Duplex Destabiliation in an Upstream Activating Sequence. J. Biol. Chem. vol. 273, no. 33, pp. 21298-21308.

[6] 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 vol. 59, pp. 3408-3426.

[7] Erzberger, J. P., Barsky, D., Scharer, O. D., Colvin, M. E., and Wilson, III, D. M. (1998). Elements in abasic site recognition by the major human and Escherichia coli apurinic/apyrimidinic endonucleases. Nucleic Acids Res. vol. 26, no. 11, pp. 2771-2778.

[8] Craigie, R. (2001). HIV Integrase, a Brief Overview from Chemistry to Therapeutics. J. of Biol. Chem. vol. 276, no. 26, pp. 23213-23216.

[9] Cantor, C. and Schimmel, P. (1980). Biophysical Chemistry. W.H. Freeman and Company, San Francisco.

[10] Bi, C. and Benham, C. J. (2004). WebSIDD: Server for predicting of the Stress-induced Duplex Destabilized (SIDD) Sites in Superhelical DNA. Bioinformatics, vol. 20, pp. 1477-1479.

[11] Benham, C. J. (2001). Stress-Induced DNA Duplex Destabilization in Transcriptional Regulation. World Scientific, Proceedings of the 2001 PacificSymposium on Biocomputing (refereed), pp. 103-114.

[12] Wang, H., Noordewier, M., and Benham, C. J. (2004). Stress-Induced DNA Duplex Destabilization (SIDD) in the Escherichia coli Genome: SIDD Sites are Closely Associated with Promoters. Genome Res. vol. 14, pp. 1575-1584.

[13] Sershen, C. L., Mell, J. C., Madden, S. M., and Benham, C. J. Superhelical Duplex Destabilization and the Recombination Position Effect. Plos One 6(6): e20798. doi:10.1371/journal.pone.0020798.

[14] Benham, C. J. (1992). Energetics of the strand separation transition in superhelical DNA, J. Mol. Biol. vol. 225, no. 835, pp. 835-847.

[15] Wang, J. C. (1979). Helical repeat of DNA in solution. Proc. Natl. Acad. Sci. vol. 76, no. 1, pp. 200-203.

[16] Glauber, R. J. (1963). Time-Dependent Statistics of the Ising Model, J. Math. Phys. vol. 4, p. 294 (1963).

[17] Privman, V. (1997). Nonequilibrium Statistical Mechanics in One Dimension. Cambridge University Press.

[18] Fehske, H., Schneider, R., and Weisse, A. (2007). Computational Many-Particle Physics. Springer.

[19] Baumgartner, A. and Binder, K. (1978). Dynamics of the Generalized Glauber-Ising Chain in a Magnetic Field. J. Stat. Phys. vol. 18, no. 5, pp. 423-444.

[20] Aranda, A., Perez-Ortin, J. E., Benham, C. J., and Del Olmo, M. L. (1997). Analysis of the Structure of a Natural Alternating d(TA)n Sequence in Yeast Chromatin. Yeast vol. 13, pp. 313-326.

[21] Sun, H., Mezel, M., Fye, R., and Benham, C. J. (1995). Monte Carlo analysis of conformational transitions in superhelical DNA. J. Chem. Phys. vol. 103, pp. 8653-65.

[22] Vologodskii, A. (2006). Simulation of equilibrium and dynamics properties of large DNA molecules, Computational Studies of DNA and RNA. Springer, pp. 579-604.

[23] Mielke, S. P., Gronbech-Jensen, N., Krishnan, V. V., Fink,W. H., and Benham, C. J. (2005). Brownian dynamics simulations of sequence-dependent duplex denaturation dynamically superhelical DNA. J. Chem. Phys., vol. 123, no. 124911.

[24] Duzdevich, D., Redding, S, and Greene, E. C. (2014). DNA dynamics and Single -Molecule Biology. Chem Rev.

[25] Baranello, L., and Kouzine. (2012). The importance of being supercoiled: How DNA mechanics regulate dynamic processes. BBA.

[26] Irobalieva, et al. (2015). Structural Diversity of supercoiled DNA. Nature Communications.