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

Chengpeng Bi[1] and Craig J. Benham[2]
UC Davis Genome Center, Universityof California, One Shields Ave., Davis CA 95616
[1]*cbi@ucdavis.edu*, [2]*cjbenham@ucdavis.edul*

Abstract

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 (SIDD), 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++.

1. 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 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 [1]. 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. Here 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.

2. Nearest neighbor energetics

Consider an N base-pair circular DNA molecule whose base sequence is $p_1p_2...p_j...p_N$,, 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

$$\alpha_r = \alpha + n/A - \mathrm{T} \qquad (1)$$

Here A is the helical repeat 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,

$$\mathrm{T} = n\tau / 2\pi \qquad (2)$$

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_r) for rotation of the single strands within denatured regions, and the residual supercoiling free energy (G_{res}).

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).

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

Here r is the number of runs. A run is a region composed entirely of separated base pairs. For the approximate method, $r \le 3$. The value $a \approx 10.16$ kcal/mol was

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

obtained by fitting to the experimental data [1] using nearest neighbor energetics.

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

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

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. 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 [2] 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}×{A, T, C, G}. The entropy also varies with the ionic concentration,

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

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,

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

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°K. R is the gas constant.

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

$$G = G_c + G_\tau + G_{res} \quad (9)$$

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

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

3. The Approximate Method

Here we use an approximate approach that finds all states whose free energy does not exceed a specified threshold [1]. 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 [1]. 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. This quantity is calculated as:

$$G(x) = \overline{G}(x) - \overline{G} \quad (12)$$

where $\overline{G}$ is the ensemble average free energy of the system and $\overline{G}(x)$ is the average free energy of all states in which the base-pair at position x is separated. The plot of $G(x)$ *versus* x is called the (helix) destabilization profile.

The algorithm was implemented in C++. The memory space takes $O(N)$, N is the DNA length. On average the running time (t) is bounded by $O(NlogN) < t \leq O(N^2)$. 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 DNA sequence. 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).

4. References

[1] C.J. Benham, "The energetics of the strand separation transition in superhelical DNA", *J. Mol. Biol.*, Elsevier, Orlando, 1992, **225**, 835-847.

[2] G. Steger, "Thermal denaturation of double-stranded nucleic acids", *Nucleic Acids Res.*, 1994, **22**, 2760-2768.