A Statistical Thermodynamic Model for Investigating the Stability of DNA Sequences from Oligonucleotides to Genomes

Garima Khandelwal[a,b], Rebecca A. Lee[b,d], B. Jayaram[a,b,c] and David L. Beveridge[d]

[a]Department of Chemistry, [b]Supercomputing Facility for Bioinformatics and Computational Biology, [c]School of Biological Sciences, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India & [d]Department of Chemistry and Molecular Biophysics Program, Wesleyan University, Middletown, CT-06459, USA

Email: garima@scfbio-iitd.res.in; Website: www.scfbio-iitd.res.in

ABSTRACT

We describe the development and testing of a simple statistical mechanics methodology for duplex DNA applicable to sequences of any composition and length and extensible to genome annotation. Following the procedure used in the COREX method for proteins (Hilser et al., *Chem. Rev*. **2006** 106:1545-1588), the microstates of a DNA sequence are modeled in terms of blocks of base pairs which are assumed to be fully closed (fully paired) or open; the block size is also a variable. This approach generates an ensemble of "bubble-like" microstates which are used to calculate the corresponding partition function. The energies of the microstates are calculated as additive contributions from hydrogen bonding, base pair stacking and solvation terms parameterized from a set of molecular dynamics simulations including solvent and ions and encompassing all possible tetranucleotide sequences (Dixit et al. *Biophys J.* **2005**, 89:3721-3740 and Lavery et al. *Nucleic Acids Research*, **2009,** 38:299-313). Thermodynamic properties and nucleotide stability constants for DNA sequences follow directly from the partition function. We tested the method by comparing computed free energies per base pair with the experimental melting temperatures of 95 oligonucleotide sequences of varying lengths and composition which yielded a correlation coefficient of -0.97. We then investigated the hypothesis that the various elements of genomes such as introns, exons and promoter regions differ in their thermodynamic stability. Plots of nucleotide stability constants vs. sequence were generated for a section of the *E. Coli* K12 genome and Granule Bound Starch Synthase I gene of *Oryza sativa*, which showed clear differentiation of the genes from promoters and introns from exons. The statistical thermodynamic model presented here, provides a new handle on the challenging problem of interpreting genomic sequences.