BIOINFORMATICS APPLICATIONS NOTE
Vol. 31 no. 3 2015, pages 421–422
doi:10.1093/bioinformatics/btu657

Genome analysis
Advance Access publication October 4, 2014

SIST: stress-induced structural transitions in superhelical DNA

Dina Zhabinskaya[1,*], Sally Madden[1] and Craig J. Benham[2,*]

[1]UC Davis Genome Center and [2]Department of Mathematics, University of California, Davis, CA 95616, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Supercoiling imposes stress on a DNA molecule that can drive susceptible sequences into alternative non-B form structures. This phenomenon occurs frequently *in vivo* and has been implicated in biological processes, such as replication, transcription, recombination and translocation. SIST is a software package that analyzes sequence-dependent structural transitions in kilobase length superhelical DNA molecules. The numerical algorithms in SIST are based on a statistical mechanical model that calculates the equilibrium probability of transition for each base pair in the domain. They are extensions of the original stress-induced duplex destabilization (SIDD) method, which analyzes stress-driven DNA strand separation. SIST also includes algorithms to analyze B-Z transitions and cruciform extrusion. The SIST pipeline has an option to use the DZCB*trans* algorithm, which analyzes the competition among these three transitions within a superhelical domain.

Availability and implementation: The package and additional documentation are freely available at https://bitbucket.org/benhamlab/sist_codes.

Contact: dzhabinskaya@ucdavis.edu

Received on August 26, 2014; revised on September 26, 2014; accepted on September 30, 2014

1 INTRODUCTION

Negative supercoiling in DNA can drive regions with susceptible sequences from the native B-form into alternate structures. Such transitions lower the superhelical energy, so they become favored at equilibrium when the stress energy they relax exceeds their cost. As the transition relaxes superhelicity, its effect is felt throughout the domain involved. This makes superhelical transitions globally competitive.

Several types of superhelical transitions have been shown to occur *in vitro*, including local strand separations, Z-DNA, cruciform extrusion and others. *In vivo*, supercoiling is imposed by DNA gyrases in prokaryotes, and by transcriptional activity in all organisms. Recently, the ssDNA-seq technique has mapped the genome-wide occurrence of *in vivo* alternate DNA structures (Kouzine *et al.*, 2013). This method found that increasing the cellular transcriptional activity substantially increases the occurrence of open regions throughout the genome. These denatured regions are accurately predicted by the SIDD algorithm.

Competing superhelical DNA structural transitions are known to play roles in important biological processes. The c-myc oncogene has a SIDD site called FUSE located 2 kb upstream from its promoters, three AluI fragments containing Z-forming sites and a potentially either G-quadriplex forming or H-forming site called the CT element 1 kb upstream from the promoters. Experiments have shown that each of these sites can be driven into their alternate structure by the superhelicity that is induced by transcription. (See Zhabinskaya and Benham, 2012) and references therein.) Cruciforms have been implicated in chromosomal translocations. (See Zhabinskaya and Benham, 2013 and references therein.)

The algorithms that underlie SIST can include any transition whose energetics have been experimentally evaluated. At present, there are only three such transitions: strand separation, B-Z transitions and cruciform extrusion. Other conformations that DNA can assume include the G-quadriplex and the H-form triplex, but their energetics are not yet fully characterized. However, the SIST algorithms are written to permit the inclusion of additional types of transitions into the model, as more experimental information about their energetics becomes available.

2 ALGORITHMS

2.1 Model

We use a statistical mechanical model to calculate the equilibrium conformational properties of a superhelical DNA. We examine individual states in which each base pair is either in B-form or in any one of the alternate conformations to which it is susceptible. If there are m conformations being considered (including the B-form) in a domain of N susceptible base pairs, there will be m^N states. Because calculations that include all states are computationally impractical, we exclude the high energy states that are not significantly occupied at equilibrium. This improves execution time while preserving the accuracy of the results. A description of the statistical mechanical model used and its algorithmic implementation have been presented elsewhere (Fye and Benham, 1999; Zhabinskaya and Benham, 2013).

The energetics of each type of transition depend strongly on the base sequence. Strand separation is easiest in Arich regions, whereas Z-form prefers CG-dinucleotide repeats and cruciform extrusion requires a high degree of inverted repeat symmetry. Deviations from these sequence preferences are energetically costly, so transitions at imperfect sites will have low probabilities at equilibrium.

2.2 Codes

The SIST software package includes two C codes, called *trans_three* and *trans_compete*. *Trans_three* allows the user to investigate each of the three types of transition (strand separation, Z-form and cruciform) individually, whereas *trans_compete* considers competitions among all three.

[*]To whom correspondence should be addressed.

To analyze cruciform extrusion, the sequence first needs to be scanned for inverted repeat sequences (IRs). Both perfect IRs and ones containing imperfections such as bulges or mismatches should be treated, and they may have a range of arm and loop lengths. We use the inverted repeats finder (IRF) algorithm for this purpose (Warburton *et al.*, 2004), downloadable from http://tandem.bu.edu/irf/irf.download.html). The parameters used in this search and the free energies that are assigned to the IRs are described in Zhabinskaya and Benham (2013). Within SIST, the *IR_finder.pl* code runs the IRF algorithm, converts the sequence into an appropriate format and assigns energies to each IR found by the IRF. It treats circular molecules differently from linear ones to assure that IRs that span the junction between the start and end of the sequence are included. The output of *IR_finder.pl* is a string of IR starts and lengths, and extrusion energies for every arm length up to full extrusion. This is necessary because extrusion can halt at any point, owing to either imperfections or full relaxation. These results are used as input to the C codes that analyze extrusion probabilities.

The *master.pl* code implements a pipeline in which the user specifies the type of analysis to be done, and the file that contains the DNA sequence. Some of the optional parameters include superhelical density (default: $\sigma = -0.06$), temperature (default: $T = 310\,\mathrm{K}$), ionic strength (default: $x = 0.01\,\mathrm{M}$) and molecular type: linear (default) or circular. The user can choose the output to list model parameters, a set of ensemble average parameter values, and the probabilities of the selected transition (or of all three transitions for *trans_compete*) for each base pair in the sequence.

As an example we use the three-way competition algorithm (*trans_compete*) to analyze an illustrative DNA sequence that was constructed by inserting three sites into the pBR322 plasmid. These are as follows: a site susceptible to strand separation (40 bp of A's), a Z-susceptible site (10 CG dinucleotides) and a perfect IR with a 4 bp loop and 22 bp arms. When we run *trans_compete* using default parameters, it provides the output probabilities shown in Figure 1. Although pbr322 contains innate melting and Z-form regions (Zhabinskaya and Benham, 2011), the three inserted regions dominate the competition at this level of superhelicity, as seen by the three peak in Figure 1.

We recommend using *trans_compete* when analyzing superhelical transitions, because studying each transition separately can overrepresent its occurrence by excluding the competition with the other types. Examples have shown that in some cases the probability of a specific transition can change from unity when studied separately to nearly zero when competition is considered (Zhabinskaya and Benham, 2013). However, studying the transitions independently might provide useful information about which regions in a genome are susceptible to each type of transition.

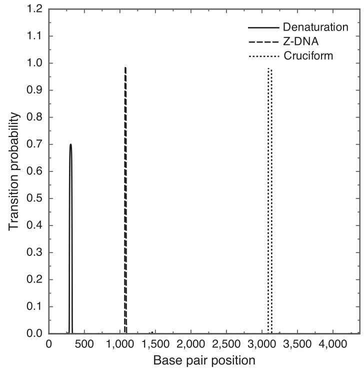

Fig. 1. Transition probabilities calculated by SIST as functions of base pair position are shown for a sequence where regions susceptible to melting, Z-DNA and cruciform extrusion were inserted into the pbr322 plasmid. The results are shown for the three-way competition algorithm (*trans_compete*) run with default parameters. The sequence and instructions on how to generate these results can be found in the example directory in the Bitbucket repository

3 DISCUSSION

The SIST package provides researchers with an accessible and flexible way to analyze the propensities of DNA sequences to form alternate structures under a variety of environmental conditions. Users can access the SIST source code through Bitbucket, which will enable them to extend the method to suit their evolving needs. They may, for example, include other transitions or update energy parameters with new experimental evidence.

In the near future, we plan also to release our windowing algorithm. This will allow users to analyze long sequences, including complete chromosomes.

ACKNOWLEDGEMENT

The authors thank Brittney Hoff for programming assistance on this project.

Funding: This research was supported by grant DBI 0850214 from the National Science Foundation.

Conflict of interest: none declared.

REFERENCES

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.

Kouzine,F. *et al.* (2013) Global regulation of promoter melting in naive lymphocytes. *Cell*, **153**, 988–999.

Warburton,P.E. *et al.* (2004) Inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.*, **14**, 1861–1869.

Zhabinskaya,D. and Benham,C.J. (2011) Theoretical analysis of the stress induced BZ transition in superhelical DNA. *PLoS Comput. Biol.*, **7**, 1–14.

Zhabinskaya,D. and Benham,C.J. (2012) Theoretical analysis of competing conformational transitions in superhelical dna. *PLoS Comput. Biol.*, **8**, 1–21.

Zhabinskaya,D. and Benham,C.J. (2013) Competitive superhelical transitions involving cruciform extrusion. *Nucleic Acids Res.*, **41**, 9610–9621.