BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2479–2481
doi:10.1093/bioinformatics/btt415

Genome analysis

Advance Access publication July 17, 2013

Twist-DNA: computing base-pair and bubble opening probabilities in genomic superhelical DNA

Daniel Jost

Laboratoire de Physique, Ecole Normale Supérieure de Lyon, CNRS UMR 5672, 69007 Lyon, France

Associate Editor: Martin Bishop

ABSTRACT

Summary: Local opening of the DNA double helix is required in many fundamental biological processes and is, in part, controlled by the degree of superhelicity imposed *in vivo* by the protein machinery. In particular, positions of superhelically destabilized regions correlate with regulatory sites along the genome. Based on a self-consistent linearization of a thermodynamic model of superhelical DNA introduced by Benham, we have developed a program that predicts the locations of these regions by efficiently computing base-pair and bubble opening probabilities in genomic DNA. The program allows visualization of results in standard genome browsers to compare DNA opening properties with other available datasets.

Availability and implementation: Source codes freely available for download at http://www.cbp.ens-lyon.fr/doku.php?id=developpement:productions:logiciels:twistdna, implemented in Fortran90 and supported on any Unix-based operating system (Linux, Mac OS X).

Contact: daniel.jost@ens-lyon.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 23, 2013; revised on July 11, 2013; accepted on July 13, 2013

1 INTRODUCTION

Many key biological processes like transcription or replication require the local opening of double-helical DNA. Under physiological conditions, most base-pairs are double-stranded, but transient melting or unwinding of *bubbles* can occur spontaneously at specific locations along the genome (Adamcik *et al*., 2012; Kowalski and Eddy, 1988).

In vivo, the double helix of DNA is mechanically constrained by protein machineries, imposing generally a negative superhelical stress to DNA with a typical superhelical density of about −0.06 in bacteria. This tendency to underwind DNA globally destabilizes the double helix, which might result in changes in the local physicochemical DNA features, like, for example, the binding constants of transcription enzymatic complexes (Schneider *et al*., 2000). It is therefore believed that highly destabilized regions in the sequence might be related to regulatory sites. Over the past 20 years, based on a thermodynamic model of superhelical DNA introduced by Craig Benham (Benham, 1992), it has been shown that these strongly *stress-induced duplex destabilized* regions significantly correlate with yeast origin of replications (Ak and Benham, 2005) or promoter regions in bacteria (Jost *et al.,* 2011; Wang *et al*., 2004). Such observed correlations suggest that information on the local opening properties of DNA could drive the identifications and annotations of putative regulatory regions in genomes.

Nine years ago, Bi and Benham developed the web server WebSIDD (Bi and Benham, 2004) that predicts the destabilization properties of genomic DNA in response to superhelical stress. However, the systematic use of this tool to genome-wide analysis of DNA local melting is highly limited by important computation times that restrict the use of the server to sequences <10 000 bp. Recently, we have developed a computationally efficient self-consistent linearization of the Benham model (Jost *et al*., 2011) that allows the prediction of position-dependent opening properties of entire genomes within minutes on a personal computer.

Here, we describe Twist-DNA, a novel freely available open source program that computes local base-pair and bubble opening probabilities of superhelical DNA, under any given temperature, ionic condition or superhelical stress density. In particular, it reads generic sequence files in FASTA format and provides output files in BED format that can be loaded into genome browsers to compare destabilization properties of genomic DNA with existing annotations or available datasets.

2 METHODS

Local opening of DNA base-pairs is described by a thermodynamic model of DNA under superhelical stress that couples the standard thermodynamic description of base-pairing with the torsional stress energetics (Benham, 1992). In this model, the free energy of a given DNA configuration is decomposed into sequence-dependent pairing and stacking free energies of closed base-pair steps, free energy penalties for the nucleation of bubbles (unpaired regions) and an effective non-local elastic-free energy accounting for torsional constraints on denatured and double-stranded DNA regions. All thermodynamic parameter values used in the model have been derived from experimental measurements (see Supplementary Table). It has to be noted that the present model neglects the writhe contribution to the superhelical density. This might lead to overestimating the opening of the double helix (Adamcik *et al*., 2012, Bar *et al*., 2012).

Resolution of the model is achieved by a self-consistent approximation that conserves the global coupling between the melting of all base-pairs, imposed by the superhelical constraint. Our approach allows the use of the efficient transfer-matrix method and the computation of position-dependent opening properties of base-pairs and bubbles along genomes. Although this approximation neglects some non-linear and finite-size effects arising from the non-local elastic free energy contribution that might be relevant for short sequences, it gives results in excellent agreement with the full resolution of the model and speeds up the computation by more than 1000-fold compared with the alternative approximate method used by WebSIDD.

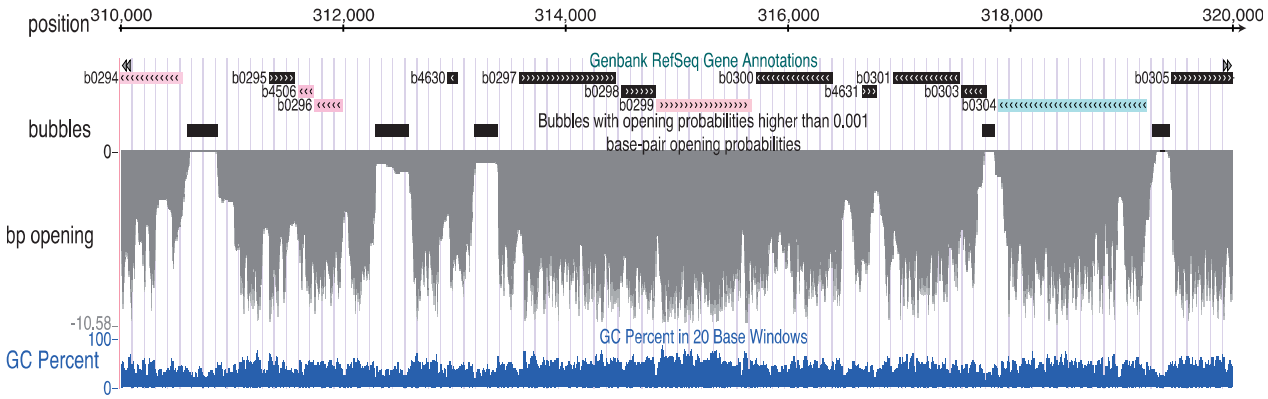

Fig. 1. Results from Twist-DNA computed for the entire genome of *E.coli* K12 (substr. MG1655) under physiological conditions, and visualized with the UCSC Microbial Genome Browser (Schneider *et al.*, 2005) for the genomic region 310–320 kbp. Base-pair opening probabilities significantly correlate with the GC-content. Bubbles are mainly localized in regulatory regions upstream of start codons

A detailed description of the model, the parameters and the computation methods can be found in Jost *et al.* (2011).

3 DESCRIPTION OF THE PROGRAM

3.1 Inputs and outputs

Twist-DNA allows the modification of several input parameters that directly influence the DNA denaturation: the superhelical density σ, the temperature T and the salt concentration $[Na^+]$.

The program reads DNA sequences in the standard FASTA format and for each position along the sequence, it computes the individual base-pair opening probability as well as the opening probabilities of bubbles of given lengths centered at this position. As superhelical stress is a global constraint, local melting properties depend on the whole input sequence. Thus, when studying a particular DNA region, the user has to be cautious about the window size taken around the investigated sequence.

As outputs, Twist-DNA produces files in BED format (with bedGraph track) that can be loaded into standard genome browsers: one containing the individual base-pair opening probabilities (in $\log_{10}$-unit) and the other containing the list of bubbles (start and end) whose opening probabilities are higher than a user-defined threshold. For example, Figure 1 shows typical stability profiles obtained for the genome of *Escherichia coli* under physiological conditions ($\sigma = -0.06$, $T = 37°C$, $[Na^+] = 0.1$ M) visualized using the UCSC microbial genome browser (Schneider *et al.*, 2005).

3.2 Implementation and performance

The source codes are written in Fortran90 and require a fortran compiler. The program has been tested to work on Unix-based operating systems (Linux and Mac OS X) using GNU compiler gfortran or Intel compiler ifort, and contains testing routines to check up its efficiency under the user's computer. Additional independent tests may be performed by the user with the profiles provided in the Supplementary Data.

The computing time linearly depends on the size of the DNA sequence and on the number of bubble sizes to investigate. For example, on a 2.66-GHz Intel Core 2 Duo PC, for the *E.coli* genome (4.6 Mbp) and 1000 bubble sizes, the program runs about 1 min to compute the base-pair opening probabilities and then runs about 15 min to estimate the bubble opening probabilities.

3.3 Example

Several bacterial genomes (including *E.coli* and *Bacillus subtilis*) were analyzed using Twist-DNA (see Supplementary Data) and compared with corresponding results for random sequences of same length and GC-content (Jost *et al.*, 2011). Genomic sequences are more destabilized with less but longer bubbles, and the DNA breathing localizes into large AT-rich regions. In particular, these metastable bubbles are mainly situated in the neighborhood upstream of transcriptional start sites and start codons (see Fig. 1) where transcription factors and RNA polymerases bind to DNA.

4 CONCLUSION

Twist-DNA is a fast and accurate program to compute base-pair opening probability profiles and the list of transiently opened bubbles along any specific genome for any user-specified conditions of temperature, salt concentration and superhelicity. Based on observed correlations between thermodynamically destabilized regions and regulatory sites, Twist-DNA can be used to discover novel properties of the regulation machineries. It can also be helpful during the design of DNA sequences or plasmids to test for example, whether specific regions are open or closed under the desired experimental condition.

ACKNOWLEDGEMENTS

The author thanks Ralf Everaers for fruitful comments on the manuscript, and Cerasela I. Calugaru for help in software development.

Conflict of Interest: none declared.

REFERENCES

Adamcik,J. *et al.* (2012) Quantifying supercoiling-induced denaturation bubbles in DNA. *Soft Matter*, **8**, 8651–8658.

Ak,P. and Benham,C.J. (2001) Susceptibility to superhelically driven DNA duplex destabilization: a highly conserved property of yeast replication origins. *PLoS Comp. Biol.*, **1**, 41–46.

Bar,A. *et al.* (2012) Denaturation of circular DNA: supercoils and overtwist. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **86**, 061904.

Benham,C.J. (1992) Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, **225**, 835–847.

Bi,C.P. and Benham,C.J. (2004) WebSIDD: server for predicting of the stress-induced duplex destabilized (SIDD) sites in superhelical DNA. *Bioinformatics*, **20**, 1477–1479.

Jost,D. *et al.* (2011) Bubble statistics and positioning in superhelically stressed DNA. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **84**, 031912.

Kowalski,D. and Eddy,M. (1988) Stable DNA unwinding, not breathing, accounts for single-stranded specific nuclease hypersensitivity of specific A+T regions. *Proc. Natl Acad. Sci. USA*, **85**, 9464–9468.

Schneider,R. *et al.* (2000) The expression of the *Escherichia coli* fis gene is strongly dependent on the superhelical density of DNA. *Mol. Microbiol.*, **38**, 167–175.

Schneider,K.L. *et al.* (2005) The UCSC archaeal genome browser. *Nucleic Acids Res.*, **34**, D407–D410.

Wang,H. *et al.* (2004) Stress-induced DNA duplex destabilization (SIDD) in the *Escherichia coli* genome: SIDD sites are closely associated with promoters. *Genome Res.*, **14**, 1575–1584.