Bioinformatics Advance Access published December 14, 2015

Genome Analysis

DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding

Tsu-Pei Chiu[1,#], Federico Comoglio[2,#], Tianyin Zhou[1,3], Lin Yang[1], Renato Paro[2] and Remo Rohs[1,*]

[1]Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA, [2]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland, [3]Present address: Google Inc., Mountain View, CA 94043, USA

[#] These authors contributed equally and are listed in alphabetic order.

*To whom correspondence should be addressed.

Associate Editor: Dr. Igor Jurisica

Abstract

Summary: DNAshapeR predicts DNA shape features in an ultra-fast, high-throughput manner from genomic sequencing data. The package takes either nucleotide sequence or genomic coordinates as input, and generates various graphical representations for visualization and further analysis. DNAshapeR further encodes DNA sequence and shape features as user-defined combinations of *k*-mer and DNA shape features. The resulting feature matrices can be readily used as input of various machine learning software packages for further modeling studies.

Availability and Implementation: The DNAshapeR software package was implemented in the statistical programming language R and is freely available through the Bioconductor project at https://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html and at the GitHub developer site https://github.com/TsuPeiChiu/DNAshapeR.git.

Contact: rohs@usc.edu

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

1 Introduction

Two distinct readout modes have emerged as crucial components of protein-DNA recognition (Abe, et al., 2015). These modes include sequence-based readout of direct contacts with the functional groups of the bases (base readout) and structure-based readout of intrinsic deviations from a canonical double helix (shape readout). DNA shape readout was originally described based on the analysis of co-crystal structures of protein-DNA complexes. Studies of DNA shape readout were then extended to massive datasets of protein-interacting DNA sequences via the use of DNAshape, a method for the high-throughput prediction of DNA structural features (Zhou, et al., 2013). Using DNAshape as the underlying tool, a motif database for transcription factor (TF) binding sites, TFBSshape (Yang, et al., 2014), and a genome browser database for DNA shape annotations, GBshape (Chiu, et al., 2014), were developed.

Rules that determine the binding affinity between TFs and their binding sites can be statistically learned from the data derived from *in vitro* high-throughput binding assays. Whereas sequence-based methods have long been used to model TF binding specificities, high-throughput prediction of DNA shape enabled us to develop methods that leverage both DNA structure and shape information. Trained with either linear regression or support vector regression algorithms, shape-augmented models were consistently shown to outperform sequence-based methods in modeling the *in vitro* binding of TFs quantitatively (Zhou, et al., 2015).

DNAshape is currently released as a stand-alone web service (Zhou, et al., 2013). Its pre-defined functionality and internet bandwidth-bounded performance made it difficult to use in genome-wide studies. To address these issues, we developed DNAshapeR, an R/Bioconductor package that can generate DNA shape predictions in an easy-to-use, easy-to-integrate and easy-to-extend manner. The output can be readily integrated into other high-throughput genomic analysis platforms.

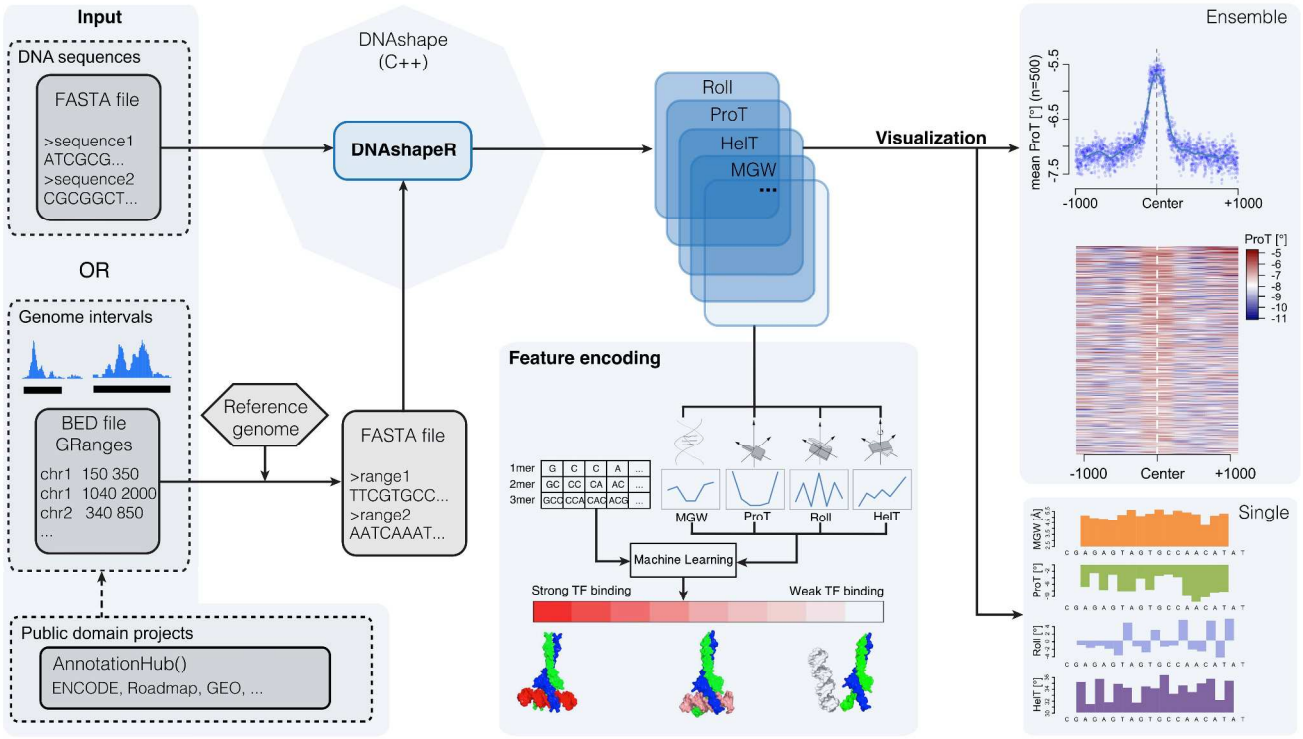

Fig. 1. Flowchart of DNAshapeR analysis. The input data can be either nucleotide sequence in FASTA file format or genomic intervals, provided by the user in BED format or derived from public databases. The core of DNAshapeR includes a high-throughput approach for the prediction of DNA shape features. MGW, HelT, ProT, and Roll can then either be visualized in the form of plots, heat maps, or genome browser tracks, or used for the assembly of feature vectors of user-defined combinations of *k*-mer and shape features.

2 High-throughput DNA Shape Prediction

The core of DNAshapeR is the DNAshape prediction method (Zhou, et al., 2013), which uses a sliding pentamer window to derive the structural features minor groove width (MGW), helix twist (HelT), propeller twist (ProT), and Roll (Fig. 1) from all-atom Monte Carlo simulations. These DNA shape features were observed in various cocrystal structures as playing an important role in specific protein-DNA binding. High-throughput predictions of DNA shape have shed light on the DNA binding specificity of TFs (He, et al., 2015; Murphy, et al., 2015) and were shown to be predictive of replication origins (Comoglio, et al., 2015).

The DNAshapeR package enables ultra-fast, high-throughput predictions of shape features for thousands of genomic sequences and generates various graphical outputs of the data (Fig. 1; Supplementary Data). The modular design of DNAshapeR enables the expansion to additional features such as conformational flexibility, biophysical properties, methylation status, to be added in future releases of the DNAshapeR package.

3 DNA Shape and *k*-mer Feature Encoding

Besides DNA shape predictions and data visualization, DNAshapeR can also be used to generate feature vectors for user-defined models. These models can consist of sequence features (1mer, 2mer, 3mer), shape features (MGW, Roll, ProT, HelT), or any combination of those features (Fig. 1; Supplementary Data). DNAshapeR encodes sequence as binary features. DNA shape features are normalized by default and can include second order shape features. The detailed definitions of sequence and shape features were provided in an earlier study (Zhou, et al., 2015).

The feature encoding function of DNAshapeR enables the generation of any user-defined subset of these features. The result of the feature encoding for each sequence is a chimera feature vector. Feature encoding of multiple sequences thus results in a feature matrix, which can be used as input for a variety of statistical machine learning methods.

Funding

This work was supported by the NIH (R01GM106056, R01HG003008 in part, U01GM103804 to R.R.). Open-source software release and open-access publication were supported by the NSF (MCB-1413539 to R.R.).

Conflict of Interest: none declared.

References

Abe, N., *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* 2015;161(2):307-318.

Chiu, T.P., *et al.* GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res* 2014.

Comoglio, F., *et al.* High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell reports* 2015;11(5):821-834.

He, Q., Johnston, J. and Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology* 2015;33(4):395-401.

Murphy, M.W., *et al.* An ancient protein-DNA interaction underlying metazoan sex determination. *Nat Struct Mol Biol* 2015;22(6):442-451.

Yang, L., *et al.* TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res* 2014;42(Database issue):D148-155.

Zhou, T., *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* 2015;112(15):4654-4659.

Zhou, T., *et al.* DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 2013;41(Web Server issue):W56-62.