Bioinformatics, 32(8), 2016, 1211–1213
doi: 10.1093/bioinformatics/btv735
Advance Access Publication Date: 14 December 2015
Applications Note

OXFORD

Genome analysis

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

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

[1]Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA, [2]Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, 4058 Basel, Switzerland and [3]Faculty of Science, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[‡]Present address: Cambridge Institute for Medical Research, Wellcome Trust/MRC Stem Cell Institute, University of Cambridge, Cambridge CB2 0XY, UK
[§]Present address: Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
Associate Editor: John Hancock

Received on 1 November 2015; revised on 8 December 2015; accepted on 9 December 2015

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, http://tsupeichiu.github.io/DNAshapeR/.
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.*, 2015), 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. Although 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 sequence 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

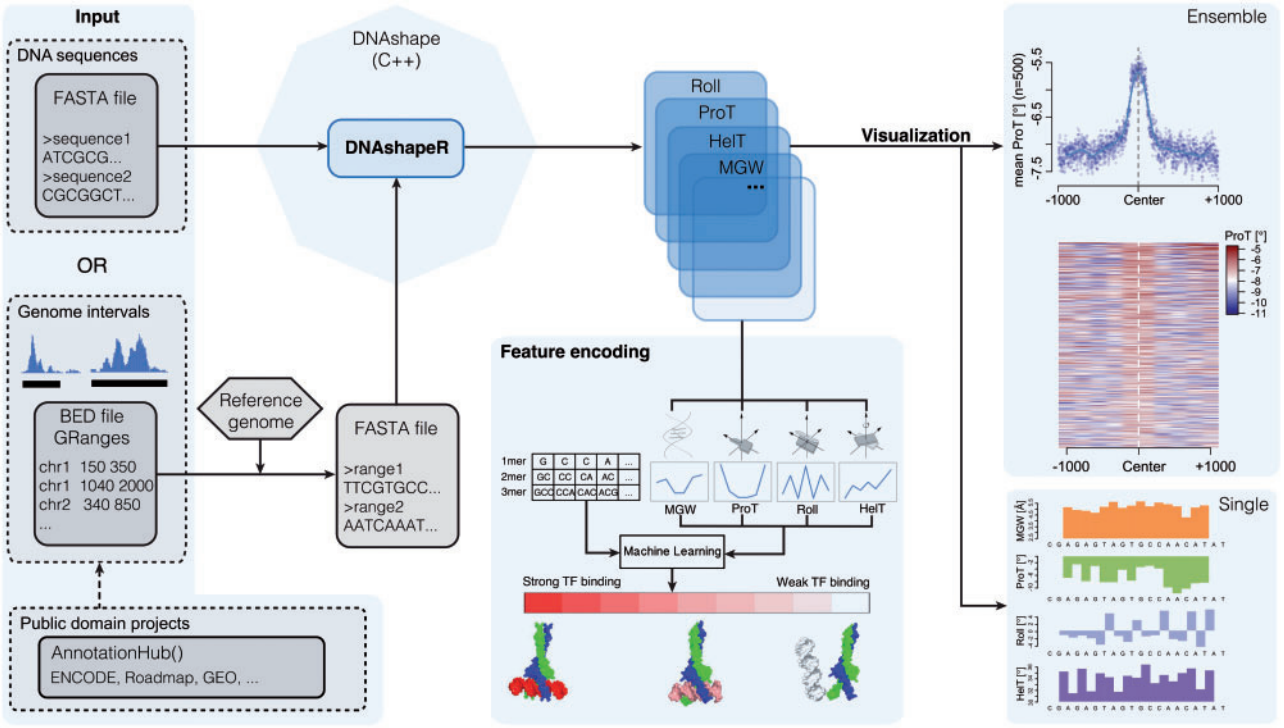

Fig. 1. Flowchart of DNAshapeR analysis. The input data can be either nucleotide sequence(s) in FASTA file format or genomic intervals, provided by the user in BED format or derived from public databases. The core of DNAshapeR includes a high-throughput approach for the prediction of DNA shape features. MGW, HelT, ProT and Roll can then 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

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.

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 achieving protein–DNA binding specificity. 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 and 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 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, and U01GM103804 to R.R.). Open-source software release and open-access publication were supported by the NSF (MCB-1413539 to R.R.). R.R. is an Alfred P. Sloan Research Fellow.

Conflict of Interest: none declared.

References

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

Chiu,T.P. *et al*. (2015) GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res*, **43**, D103–D109.

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

He,Q. *et al*. (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol*., **33**, 395–401.

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

Yang,L. *et al*. (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res*., **42**(Database issue), D148–D155.

Zhou,T. *et al*. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res*., **41** (Web Server issue), W56–W62.

Zhou,T. *et al*. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. USA*, **112**, 4654–4659.