Package 'DNAshapeR'

April 27, 2025

Type Package

Title High-throughput prediction of DNA shape features

Version 1.36.0

Date 2018-06-08

Author Tsu-Pei Chiu and Federico Comoglio

Maintainer Tsu-Pei Chiu <tsupeich@usc.edu>

Description DNAhapeR is an R/BioConductor package for ultra-fast,
high-throughput predictions of DNA shape features.
The package allows to predict,
visualize and encode DNA shape features
for statistical learning.

License GPL-2

biocViews StructuralPrediction, DNA3DStructure, Software

LazyData TRUE

Depends R (>= 3.4), GenomicRanges

Imports Rcpp (>= 0.12.1), Biostrings, fields

LinkingTo Rcpp

NeedsCompilation yes

VignetteBuilder knitr

Suggests AnnotationHub, knitr, rmarkdown, testthat,
BSgenome.Scerevisiae.UCSC.sacCer3, BSgenome.Hsapiens.UCSC.hg19,
caret

RoxygenNote 6.0.1

git_url https://git.bioconductor.org/packages/DNAshapeR

git_branch RELEASE_3_21

git_last_commit b6ee88f

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-04-27

Contents

`DNAshapeR-package` *DNAshapeR package for high-throughput prediction of DNA shape features*

Description

The main functions in the package are `getFasta` and `getShape`. Shape predictions can be additionally plotted with `plotShape`. See package vignette for examples.

Details

All support questions should be posted to the Bioconductor support site: `support.bioconductor.org`, using DNAshapeR as a tag.

Author(s)

Tsu-Pei Chiu and Federico Comoglio

References

DNAshapeR reference:

T.-P. Chiu*, F. Comoglio*, T. Zhou, L. Yang, R. Paro, and R. Rohs: DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding (2016). Bioinformatics `http://bioinformatics.oxfordjournals.org/content/early/2016/01/09/bioinformatics.btv735` (*equal contributor in alphabetic order)

convertMethFile *Convert fasta file to methylated file format*

Description

Convert fasta file to methylated file format

Usage

```
convertMethFile(fastaFileName, methPositionFileName)
```

Arguments

fastaFileName	The name of the input fasta format file, including full path to file if it is located outside the current working directory.
methPositionFileName	The name of the input position file indicating the methlation position

Value

methFileName fasta file containing methylated Cytosine

Author(s)

Satyanarayan Rao & Tsu-Pei Chiu

encodeKmerHbond *encode Hbond*

Description

encode Hbond

Usage

```
encodeKmerHbond (k, dnaStringSet )
```

Arguments

k	k-mer sequence
dnaStringSet	dnaStringSet

Value

featureVector

encodeKMerSeq *Encode k-mer DNA sequence features*

Description

DNAshapeR can be used to generate feature vectors for a user-defined model. The model can be a k-mer sequence. Sequence is encoded in four binary features (i.e., in terms of 1-mers, 0001 for adenine, 0010 for cytosine, 0100 for guanine, and 1000 for thymine) at each nucleotide position (Zhou, et al., 2015). The function permits an encoding of 2-mers and 3-mers (16 and 64 binary features at each position, respectively).

Usage

```
encodeKMerSeq(k, dnaStringSet)
```

Arguments

k	A number indicating k-mer sequence encoding
dnaStringSet	A DNAStringSet object of the inputted fasta file

Value

featureVector A matrix containing encoded features. Sequence feature is represented as binary numbers

Author(s)

Tsu-Pei Chiu

encodeNstOrderShape *Encode n-st order shape features DNAshapeR can be used to generate feature vectors for a user-defined model. The model can be a shape model. There are four structural parameters including MGW, Roll, ProT and HelT. The second order shape features are product terms of values for the same category of shape features at adjacent positions.*

Description

Encode n-st order shape features DNAshapeR can be used to generate feature vectors for a user-defined model. The model can be a shape model. There are four structural parameters including MGW, Roll, ProT and HelT. The second order shape features are product terms of values for the same category of shape features at adjacent positions.

Usage

```
encodeNstOrderShape(n, shapeMatrix, shapeType)
```

Arguments

n	A number indicating n-st order shape encoding
shapeMatrix	A matrix containing DNAshape prediction result
shapeType	A character name of shape (MGW, Roll, ProT, HelT) features

Value

featureVector A matrix containing encoded features. shape feature is represented as continuous numbers

Author(s)

Tsu-Pei Chiu

---

encodeSeqShape *Encode k-mer DNA sequence and n-th order DNA Shape features*

Description

DNAshapeR can be used to generate feature vectors for a user-defined model. These models can be based on DNA sequence (1-mer, 2-mer, 3-mer) or DNA shape (MGW, Roll, ProT, HelT) features or any combination thereof. Sequence is encoded as four binary features (i.e., 0001 for adenine, 0010 for cytosine, 0100 for guanine, and 1000 for thymine, for encoding of 1-mers) at each nucleotide position (Zhou, et al., 2015). Encoding of 2-mers and 3-mers (16 and 64 binary features at each position, respectively) is also supported. Shape features include first and second order (or higher order) values for the four structural parameters MGW, Roll, ProT and HelT. The second order shape features are product terms of values for the same category of shape features at adjacent positions. The function allows to generate any subset of these features, e.g. a given shape category or first order shape features, and any desired combination of shape and sequence features. Feature encoding returns a feature matrix for a dataset of multiple sequences, in which each sequence generates a concatenated feature vector. The output of this function can be used directly for any statistical machine learning method.

Usage

```
encodeSeqShape(fastaFileName, shapeMatrix, featureNames, normalize)
```

Arguments

fastaFileName	A character name of the input fasta format file, including full path to file if it is located outside the current working directory.
shapeMatrix	A matrix containing DNAshape prediction result
featureNames	A vector containing a combination of user-defined sequence and shape parameters. The parameters can be any combination of "k-mer", "n-shape", "n-MGW", "n-ProT", "n-Roll", "n-HelT" (k, n are integers)
normalize	A logical indicating whether to perform normalization. Default to TRUE.

Value

featureVector A matrix containing encoded features. Sequence features are represented as binary numbers, while shape features are represented as real numbers.

Author(s)

Tsu-Pei Chiu

Examples

```
fn <- system.file("extdata", "CGRsample_short.fa", package = "DNAshapeR")
pred <- getShape(fn)
featureNames <- c("1-shape")
featureVector <- encodeSeqShape(fn, pred, featureNames)
```

---

getFasta *Extract fasta sequence given a set of genomic intervals and a reference genome.*

Description

DNAshapeR can predict DNA shape features from custom FASTA files or directly from genomic coordinates in the form of a GRanges object within BioConductor (see <https://bioconductor.org/packages/release/bioc/html/
for more information).

Usage

```
getFasta(GR, BSgenome, width = 1e3, filename = 'tmp.fa')
```

Arguments

GR	A GRanges object indicating genomic coordinates
BSgenome	A BSgenome object indicating the genome of interest
width	A number indicating a fixed width of sequences
filename	The Name of the input fasta format file, including full path to file if it is located outside the current working directory

Value

writes a fasta file

Author(s)

Federico Comoglio

Examples

```
gr <- GRanges(seqnames = c("chrI"),
strand = c("+", "-", "+"),
ranges = IRanges(start = c(100, 200, 300), width = 100))
library(BSgenome.Scerevisiae.UCSC.sacCer3)
getFasta(gr, BSgenome = Scerevisiae, width = 100, filename = "tmp.fa")
fn <- "tmp.fa"
pred <- getShape(fn)
```

`getShape` *Predict DNA shape from a FASTA file*

Description

The DNA prediction uses a sliding pentamer window where structural features unique to each of the 512 distinct pentamers define a vector of minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT) at each nucleotide position (Zhou, et al., 2013). MGW and ProT define base-pair parameter whereas Roll and HelT represent base pair-step parameters. The values for each DNA shape feature as function of its pentamer sequence were derived from all-atom Monte Carlo simulations where DNA structure is sampled in collective and internal degrees of freedom in combination with explicit counter ions (Zhang, et al., 2014). The Monte Carlo simulations were analyzed with a modified Curves approach (Zhou, et al., 2013). Through data mining, average values for each shape feature were calculated for the on average 44 occurrences of each pentamer in an ensemble of Monte Carlo trajectories for 2,121 DNA fragments of 12-27 base pairs in length. DNAshapeR predicts four DNA shape features, which were observed in various co-crystal structures playing an important role in specific protein-DNA binding. The core prediction algorithm enables ultra-fast, high-throughput predictions of shape features for thousands of genomic sequences and is implemented in C++. Since it is likely that features describing additional structural properties or equivalent features derived from different experimental or computational sources will become available, the package has a flexible modular design that easily allows future expansions. In the latest version, we further added additional 9 DNA shape features beyond our previous set of 4 features, and expanded our available repertoire to a total of 13 features, including 6 inter-base pair or base pair-step parameters (HelT, Rise, Roll, Shift, Slide, and Tilt), 6 intra-base pair or base pair-step parameters (Buckle, Opening, ProT, Shear, Stagger, and Stretch), and MGW.

Usage

```
getShape(filename, shapeType = 'Default', parse = TRUE,
methylate = FALSE, methylatedPosFile = NULL)
```

Arguments

`filename`	The name of input fasta format file, including full path to file if it is located outside the current working directory.
`shapeType`	A character indicating the shape parameters which can be "MGW", "ProT", "Roll", "HelT" or "All" (meaning all four shapes)
`parse`	A logical value indicating whether parse the prediction result

`methylate` A logical value indicating wheter consider methlatation

`methylatedPosFile`
The name of input postion file indicating methlated position

Details

Predict biophysical feature

Our previous work explained protein-DNA binding specificity based on correlations between MGW and electrostatic potential (EP) observed in experimentally available structures (Joshi, et al., 2007). However, A/T and C/G base pairs carry different partial charge distributions in the minor groove (due primarily to the guanine amino group), which will affect minor-groove EP. We developed a high-throughput method to predict minor-groove EP based on data mining of results from solving the nonlinear Poisson-Boltzmann calculations (Honig & Nicholls, 1995) on 2,297 DNA structures derived from Monte Carlo simulations. DNAshapeR includes EP as an additional feature.

Value

shapeList A List containing shapre prediction result

Author(s)

Federico Comoglio & Tsu-Pei Chiu

Examples

```
fn <- system.file("extdata", "CGRsample.fa", package = "DNAshapeR")
pred <- getShape(fn)
```

`heatShape` *Plot heatmap of DNA shape features*

Description

Plot heatmap of DNA shape features

Usage

```
heatShape(shapeMatrix, nBins, ordRow = NULL, useRaster = TRUE, ... )
```

Arguments

`shapeMatrix`	A matrix containing DNAshape prediction results.
`nBins`	An integer specifying the number of equally-sized bins in which shape predictions should be aggregated. Summarized predictions can be visualized by setting nBins=1.
`ordRow`	A numeric vector (of the same length as the number of rows of shapeMatrix) defining the permutation of the rows of shapeMatrix to be used for plotting. Default to NULL, i.e. rows are ordered by coefficients of variation.

useRaster	Logical, if TRUE a bitmap raster is used to plot the image instead of polygons (see ?graphics::image for details).
...	Additional parameters to be passed to the image.plot function (see ?fields::image.plot for details).

Value

Called for its effects

Author(s)

Federico Comoglio

Examples

```
fn <- system.file("extdata", "CGRsample.fa", package = "DNAshapeR")
pred <- getShape(fn)
library(fields)
heatShape(pred$MGW, 20)
```

`normalize` *Min-Max normalization*

Description

Min-Max normalization

Usage

```
normalize(x, max, min)
```

Arguments

x	A matrix containing encoded features
max	A number maximum number for Min-Max Normalization
min	A number minimum number for Min-Max Normalization

Value

featureVector A matrix containing encoded features. shape feature is represented as continuous numbers

Author(s)

Tsu-Pei Chiu

normalizeShape *Normalize n-st order shape features*

Description

Normalize n-st order shape features

Usage

```
normalizeShape(featureVector, thOrder, shapeType, normalize)
```

Arguments

featureVector	A matrix containing encoded features.
thOrder	A number indicating n-st order shape encoding
shapeType	A character name of shape (MGW, Roll, ProT, HelT) features
normalize	A logical indicating whether to perform normalization. Default to TRUE.

Value

featureVector A matrix containing encoded features.

Author(s)

Tsu-Pei Chiu

plotShape *Plot metaprofiles of DNA shape features*

Description

DNA shape features can be visualized as aggregated line plots (also known as metaprofiles, see Comoglio et al., 2015), heat maps (Yang et al., 2014) and genome browser tracks (Chiu et al., 2014).

Usage

```
plotShape(shapeMatrix, background = NULL,
colDots = rgb( 0, 0, 1, 0.1),
colDotsBg = rgb( 0, 0, 0, 0.1),
colLine = 'steelblue', colLineBg = 'gray50', cex = 0.5, lwd = 2, ylim, ...)
```

Arguments

shapeMatrix	A matrix containing DNAshape prediction results
background	A matrix containing DNAshape prediction results for a set of background regions. Default to NULL, i.e. background not provided.
colDots	A character vector specifying the color of the points representing the column mean of shapeMatrix. Default to rgb(0, 0, 1, 0.1).
colDotsBg	A character vector specifying the color of the points representing the column mean of background. Default to rgb(0, 0, 0, 0.1).
colLine	A character string giving the color name of line representing the column mean of shapeMatrix. Default to 'steelblue'.
colLineBg	A character string giving the color name of line representing the column mean of background. Default to 'gray50'.
cex	A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default. Default to 0.5.
lwd	A numerical value specifying the line width. Default to 2.
ylim	A numerical vector of size 2 specifying the y-axis plot range.
...	Additional parameters to be passed to the R plot function.

Value

Called for its effects

Author(s)

Federico Comoglio

Examples

```
fn <- system.file("extdata", "CGRsample.fa", package = "DNAshapeR")
pred <- getShape(fn)
plotShape(pred$MGW)
plotShape(pred$ProT)
plotShape(pred$Roll)
plotShape(pred$HelT)
```

---

readNonStandardFastaFile

Read the position fasta file

Description

Read the position fasta file

Usage

```
readNonStandardFastaFile(filename)
```

Arguments

filename	The name of the input position file indicating the methlation position

Value

df dataframe

Author(s)

Satyanarayan Rao & Tsu-Pei Chiu

---

readShape *Read (parse) DNA shape predictions*

Description

Read DNA shape predictions

Usage

```
readShape(filename)
```

Arguments

filename	character name of the file containing shape predictions, including full path to file if it is located outside the current working directory.

Value

shapeMatrix matrix containing the shape prediction result

Author(s)

Federico Comoglio & Tsu-Pei Chiu

Examples

```
fn <- system.file("extdata", "CGRsample.fa", package = "DNAshapeR")
pred <- readShape(fn)
```

trackShape *Plot track view of DNA shape features*

Description

Plot track view of DNA shape features

Usage

```
trackShape( filename, shapeList )
```

Arguments

filename	The name of the input fasta format file, including full path to file if it is located outside the current working directory
shapeList	A list containing four DNAshape prediction results

Value

Called for its effects

Note

None.

Author(s)

Tsu-Pei Chiu

Examples

```
fn2 <- system.file("extdata", "SingleSeqsample.fa", package = "DNAshapeR")
pred2 <- getShape(fn2)
trackShape(fn2, pred2) # Only for single sequence file
```

Index