**Supplementary Data – Bioconductor Vignette for DNAshapeR Package**

# 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, [3]Faculty of Science, University of Basel, Klingelbergstrasse 50, 4046 Basel, Switzerland

[&]Present address: Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

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

*To whom correspondence should be addressed: rohs@usc.edu

## Introduction

DNAshapeR predicts DNA shape features in an ultra-fast, high-throughput manner from genomic sequencing data. The package takes either nucleotide sequence or genomic intervals as input, and generates various graphical representations for further analysis. DNAshapeR further encodes DNA sequence and shape features for statistical learning applications by concatenating feature matrices with user-defined combinations of *k*-mer and DNA shape features that can be readily used as input for machine learning algorithms.

In this vignette, you will learn:

- how to load/install DNAshapeR
- how to predict DNA shape features
- how to visualize DNA shape predictions
- how to encode sequence and shape features, and apply them

## Load DNAshapeR

```
library(DNAshapeR)
```

## Predict DNA shape features

The core of DNAshapeR, the DNAshape method (Zhou, et al., 2013), 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. MGW and ProT define base-pair parameters 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 sodium counter ions (Zhang, et al., 2014). The Monte Carlo simulations

were analyzed with a modified Curves approach (Zhou, et al., 2013). Average values of each shape feature for each pentamer were derived from analyzing the ensemble of Monte Carlo predictions for 2,121 DNA fragments of 12−27 base pairs in length. DNAshapeR predicts the four DNA shape features MGW, HelT, ProT, and Roll, which were observed in various cocrystal structures as playing an important role in specific protein-DNA binding.

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.

## 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, named DNAphi, 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.

## Predict DNA shape feature due to CpG methylation

To achieve a better mechanistic understanding of the effect of CpG methylation on local DNA structure, we developed a high-throughput method, named *methyl*-DNAshape, for predicting the impact of cytosine methylation on DNA shape features. In analogy to the DNAshape method (Zhou, et al., 2013), the method predicts DNA shape features (ProT, HelT, Roll, and MGW) in the context of CpG methylation based on methyl-DNAshape Pentamer Query Table (mPQT) derived from the results of all-atom Monte Carlo simulations on a total of 3,518 DNA fragments of lengths varying from 13 to 24 bp.

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/GenomicRanges.html for more information).

### From FASTA file

To predict DNA shape features from a FASTA file

```
library(DNAshapeR)

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

pred <- getShape(fn)
```

```
## Reading the input sequence......

## Reading the input sequence......

## Reading the input sequence......

## Reading the input sequence......

## Parsing files......

## Record length: 2000

## Record length: 1999

## Record length: 2000

## Record length: 1999

## Done
```

## From genomic intervals (e.g. TFs binding sites, CpG islands, replication origins, …)

To predict DNA shape from genomic intervals stored as GRanges object, a reference genome is required. Several reference genomes are available within BioConductor as BSgenome objects (see http://bioconductor.org/packages/release/bioc/html/BSgenome.html for more information). For example, the sacCer3 release of the *S.Cerevisiae* genome can be retrieved by

```
# Install Bioconductor packages

source("http://bioconductor.org/biocLite.R")

biocLite("BSgenome.Scerevisiae.UCSC.sacCer3")



library(BSgenome.Scerevisiae.UCSC.sacCer3)
```

Given a reference genome, the **getFasta** function first extracts the DNA sequences based on the provided genomic coordinates, and then performs shape predictions within a user-defined window (of size equal to width, 100 bp in the example below) computed from the center of each genomic interval:

```
# Create a query GRanges object

gr <- GRanges(seqnames = c("chrI"),

              strand = c("+", "-", "+"),

              ranges = IRanges(start = c(100, 200, 300), width = 100))

getFasta(gr, Scerevisiae, width = 100, filename = "tmp.fa")

fn <- "tmp.fa"

pred <- getShape(fn)
```

## From public domain projects

The genomic intervals can also be obtained from public domain projects, including ENCODE, NCBI, Ensembl, etc. The AnnotationHub package
(see http://bioconductor.org/packages/release/bioc/html/AnnotationHub.html for more information) provides an interface to retrieve genomic intervals from these multiple online project resources.

```
# Install Bioconductor packages

library(BSgenome.Hsapiens.UCSC.hg19)

library(AnnotationHub)
```

The genomic intervals of interest can be selected progressively through the functions of **subset** and **query** with keywords, and can be subjected as an input of GRanges object to **getFasta** function.

```
ah <- AnnotationHub()

ah <- subset(ah, species=="Homo sapiens")

ah <- query(ah, c("H3K4me3", "Gm12878", "Roadmap"))

getFasta(ah[[1]], Hsapiens, width = 150, filename = "tmp.fa")

fn <- "tmp.fa"

pred <- getShape(fn)
```

## From FASTA file with methylated DNA sequence

To predict DNA shape features in the context of CpG methylation, one can prepare a FASTA file of sequence with symbol ‘Mg’: ‘M’ referring to cytosine of methylated CpG on the leading strand and ‘g’ referring the cytosine of methylated CpG on lagging strand. For example,

> seq1

GTGTCACMgCGTCTATACG

notifying the cytosine at position 8$^{th}$ on the leading strand and the one at position 9$^{th}$ on the lagging strand are methylated.

```
library(DNAshapeR)

fn <- system.file("extdata", "MethylSample.fa", package = "DNAshapeR")

pred <- getShape(fn, methylate = TRUE)
```

## From FASTA and methylated position files

To predict DNA shape features in the context of CpG methylation, in addition to providing regular FASTA file (without symbolizing ‘Mg’) one can provide an additional input file identifying methylated positions. For example,

> seq1

4,16

notifying the cytosine at position 4$^{th}$ and 16$^{th}$ on leading strand, and 5$^{th}$ and 17$^{th}$ on lagging strand are methylated.

```
library(DNAshapeR)

fn <- system.file("extdata", "SingleSeqsample.fa", package = "DNAshapeR")

fn_pos <- system.file("extdata", "MethylSamplePos.fa", package = "DNAshapeR")

pred <- getShape(fn, methylate = TRUE, methylatedPosFile = fn_pos)
```

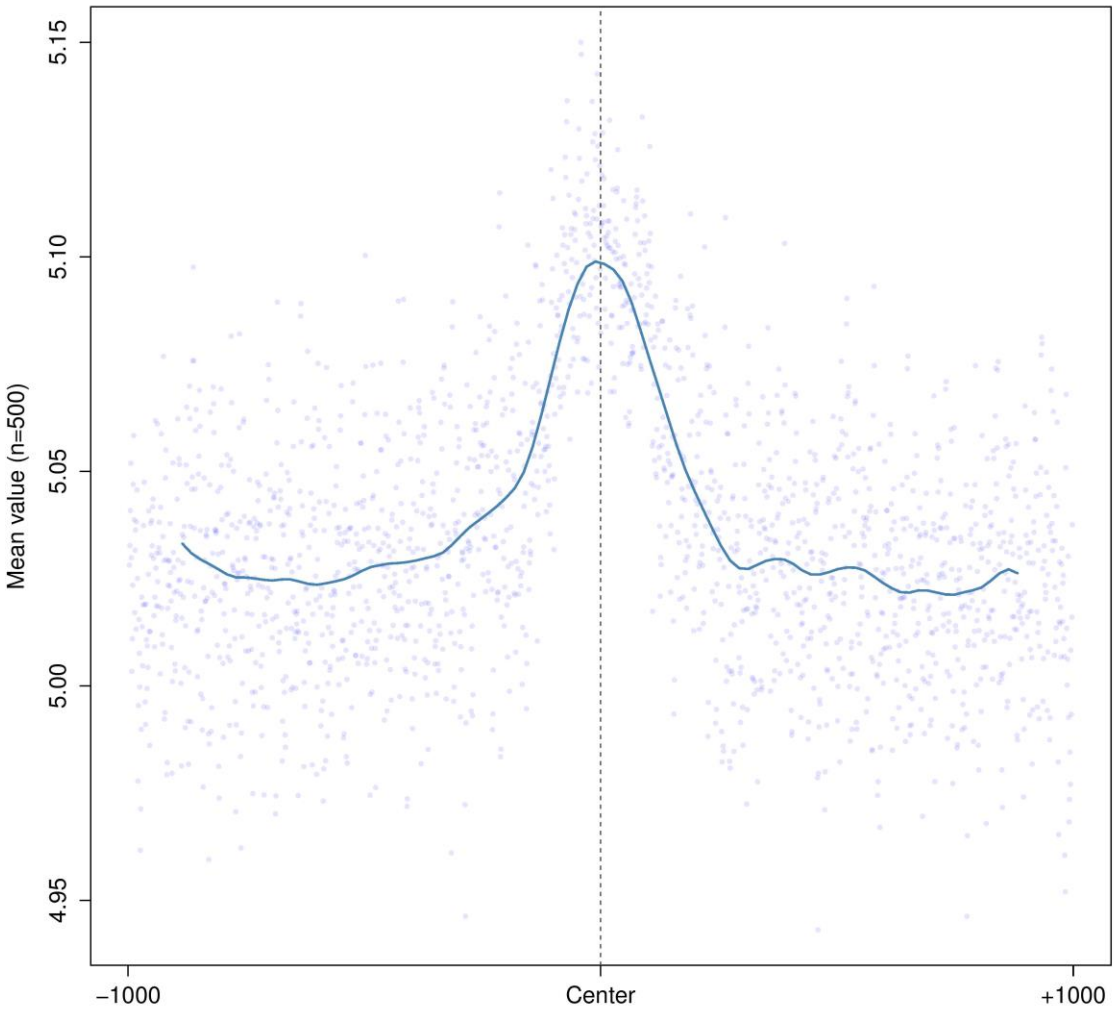
## Visualize DNA shape prediction

DNAshapeR can be used to generate various graphical representations for further analyses. The prediction result can be visualized in the form of scatter plots (as introduced in Comoglio, et al., 2015), heat maps (as introduced in Yang, et al., 2014), or genome browser tracks (as introduced in Chiu, et al., 2014).

### Ensemble representation: metashape plot

The prediction result can be visualized in the metaprofiles of DNA shape features.

```
plotShape(pred$MGW)
```



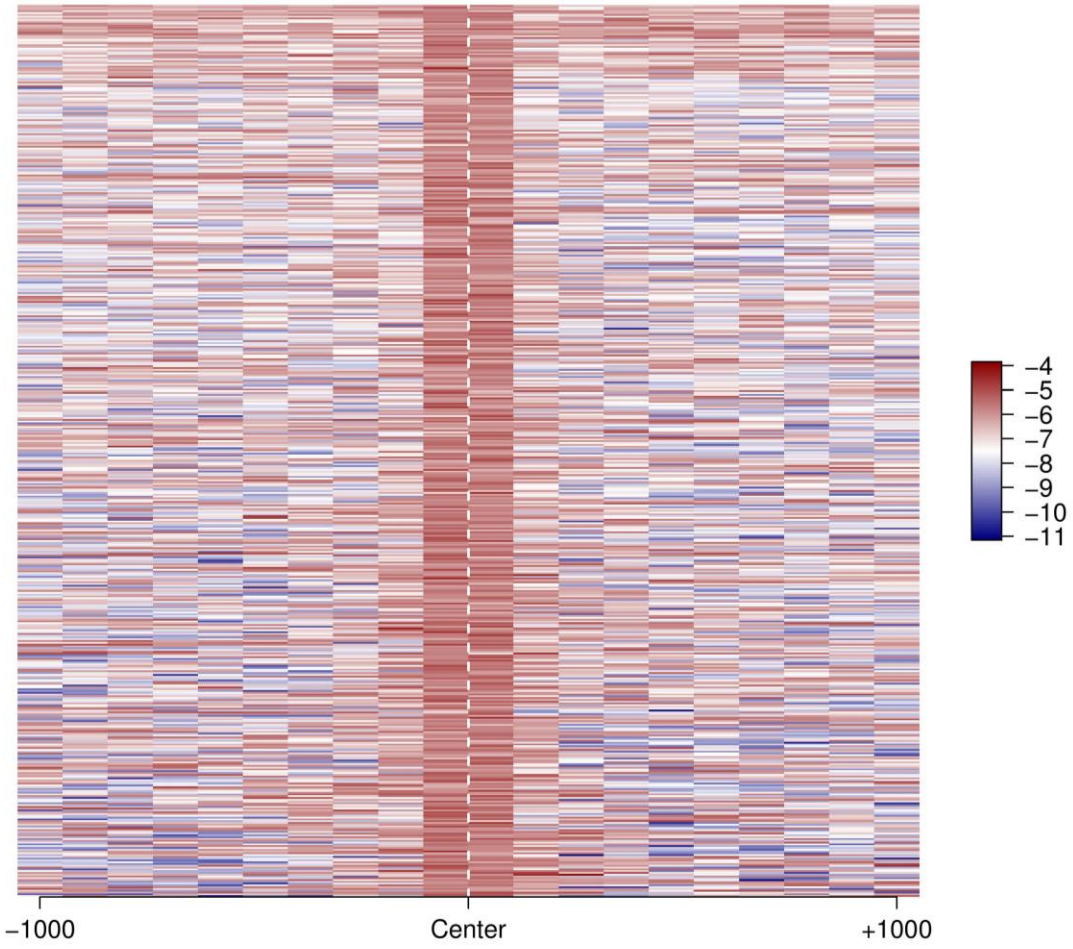

```
# MGW can be replaced with HelT, Rise, Roll, Shift, Slide, Tilt, Buckle, Opening, ProT, Shear, Stagger, Stretch or EP
```

## Ensemble representation: heat map

The prediction result can be visualized in the heat map of DNA shape features.

```
library(fields)

heatShape(pred$ProT, 20)
```




```
# ProT can be replaced with MGW can be replaced with MGW, Buckle, Opening, ProT, Shear,
Stagger, Stretch or EP

#heatShape(pred$Roll[1:500, 1:1980], 20)

# Roll can be replaced with HelT, Rise, Roll, Shift, Slide or Tilt
```

## Individual representation: genome browser-like tracks

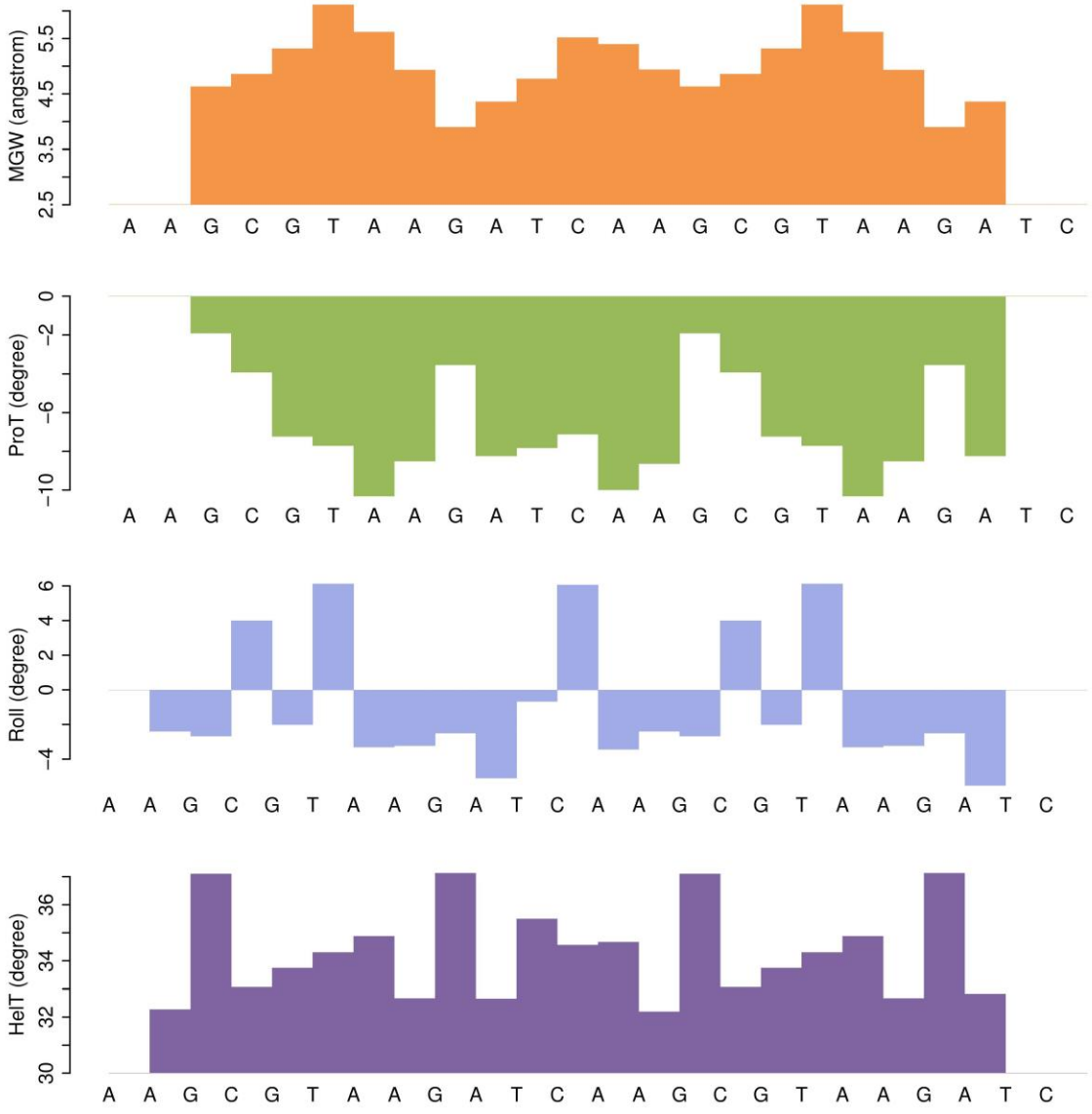
The prediction result can be visualized in the form of genome browser tracks.

*Note that the input data should only contain one sequence.

```
fn2 <- system.file("extdata", "SingleSeqsample.fa", package = "DNAshapeR")

pred2 <- getShape(fn2)

trackShape(fn2, pred2) # Only for single sequence files
```




## Encode sequence and shape features

DNAshapeR can be used to generate feature vectors for a user-defined model. These models can consist of either sequence features (1-mer, 2-mer, 3-mer), shape features (MGW, HelT, Rise, Roll, Shift, Slide, Tilt, Buckle, Opening, ProT, Shear, Stagger, Stretch and EP), or any combination of those two. For 1-mer features, sequence is encoded in form of four binary numbers (i.e., in terms of 1-mers, 1000 for adenine, 0100 for cytosine, 0010 for guanine, and 0001 for thymine) at each nucleotide position (Zhou, et al., 2015). The feature encoding function of the DNAshapeR package enables the

determination of higher order sequence features, for example, 2-mers and 3-mers (16 and 64 binary features at each position, respectively).

User can also choose to include second order shape features in the generated feature vector. The second order shape features are product terms of values for the same category of shape features (MGW, HelT, Rise, Roll, Shift, Slide, Tilt, Buckle, Opening, ProT, Shear, Stagger, Stretch or EP) at adjacent positions. They were introduced to encode the tendency of, for instance, a narrow minor groove region exhibiting an enhanced narrowing if adjacent positions are also characterized by a narrow groove (Zhou, et al., 2015). The feature encoding function of DNAshapeR enables the generation of any subset of these features, either only a selected shape category or first order shape features, and any combination with shape or sequence features. The result of feature encoding for each sequence is a chimera feature vector.

## Encoding process

A feature type vector should be defined before encoding. The vector can be any combination of characters of “k-mer”, “n-shape”, “n-MGW”, “n-ProT”, “n-Roll”, “n-HelT”, “n-Rise”, “n-Shift”, “n-Slide”, “n-Tilt”, “n-Buckle”, “n-Opening”, “n-Shear”, “n-Stagger”, “n-Stretch”, “n-EP” (k, n are integers) where “1-shape” refers to first order and “2-shape” to second order shape features. Notice that n-shape represents the default 4 shape features (MGW, ProT, Roll and HelT).

```
library(Biostrings)

featureType <- c("1-mer", "1-shape")

featureVector <- encodeSeqShape(fn, pred, featureType)

featureVector
```

## Showcase of statistical machine learning application

Feature encoding of multiple sequences thus results in a feature matrix, which can be used as input for variety of statistical machine learning methods. For example, an application is the quantitative modeling of PBM derived protein-DNA binding by linear regression as demonstrated below.

First, the experimental binding affinity values are combined with the feature matrix in a data frame structure.

```
filename3 <- system.file("extdata", "PBMsample.s", package = "DNAshapeR")

experimentalData <- read.table(filename3)

df <- data.frame(affinity=experimentalData$V1, featureVector)
```

Then, a machine learning package (which can be any learning tools) is used to train a multiple linear regression (MLR) model based on 10-fold cross-validation. In this example, we used the caret package (see http://caret.r-forge.r-project.org/ for more information).

```
library(caret)



trainControl <- trainControl(method = "cv", number = 10, savePredictions = TRUE)

model <- train (affinity~ ., data = df, trControl=trainControl, method="lm", preProcess=NULL)

summary(model)
```

## Author contributions

T.P.C., F.C., and R.R. designed and executed this project with the help of T.Z., L.Y., and R.P. based on methods developed by T.P.C., T.Z., and L.Y. J.L. introduced additional DNA shape features. The project was conceived by F.C. and directed by R.R.

## Acknowledgements

The authors acknowledge comments and suggestions by members of the Rohs lab. This work was supported by the NIH (R01GM106056, U01GM103804, and R01HG003008 to R.R.). Release of open-source software and open-access publication were supported by the NSF (MCB-1413539 to R.R.).

## References

Chiu, T.-P., et al. GBshape: a genome browser database for DNA shape annotations. Nucleic Acids Res. 2015;43:D103-109.

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

Yang, L., et al. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014;42: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:W56-62.

Zhang, X., et al. Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. Nucleic Acids Res. 2014;42(4):2789-2797.

Honig, B. and Nicholls, A. Classical electrostatics in biology and chemistry. Science, 1995;268: 1144-1149.

Joshi, R., et al. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell, 2007;131:530-543.