DiffLogo user guide

Hendrik Treutler

October 17, 2016

1 Introduction

The DiffLogo tool is a R package for the visualization of differences between multiple motifs for different alphabets. The user supplies a set of motifs each represented as position weight matrices (PWMs) [1]. The DiffLogo package supports the comparison of two motifs by a single difference logo and the comparison of multiple motifs by a table of difference logos. Difference logos are based on the idea behind the well-known sequence logo [2], i.e. motifs are visualized position-wise based on two functions. First, the *stackHeight* function computes the height of each stack. Second, the *baseDistribution* function breaks down the stack height on the individual characters. The user is able to parametrise the individual functions with arbitrary functions *stackHeight* and *baseDistribution*. Default implementations of both functions are provided.

2 Download and import library

After installing the package, the user is able to import DiffLogo.

```
> library(DiffLogo)
```

3 Import PWMs

PWMs can be represented as object of type pwm, data.frame, or matrix. The user is able to import motifs from any source in one of these formats.

```
> library(MotifDb)
> ## import motifs
> hitIndeces <- grep ('CTCF',
+                     values (MotifDb)$geneSymbol,
+                     ignore.case=TRUE)
> list <- as.list(MotifDb[hitIndeces])
> ## get two motifs
> pwm1 <- list$"Hsapiens-jolma2013-CTCF"
> ## trim and reverse complement
> pwm2 <- list$"Hsapiens-JASPAR_CORE-CTCF-MA0139.1"[4:1, 18:2]
```

Here, we import two motifs from the transcription factor CTCF from package *MotifDb* [3]. Alternatively, there are example PWMs in folder *extdata/pwm* and *extdata/alignments* shipped with the package DiffLogo. (CTCF motifs extracted from [4], E-Box transcription factor binding sites extracted from [5], and F-Box protein domains extracted from [6]).

```
> ## import nine DNA motifs for transcription factor CTCF from matrix
> motif_folder <- "extdata/pwm"
> motif_names_dna = c(
+   "GM12878", "H1-hESC", "HeLa-S3", "HepG2",   "HUVEC",
+   "K562",    "MCF7",    "NHEK",    "ProgFib")
> motifs_dna = list()
> for (name in motif_names_dna) {
+   fileName <- paste(motif_folder,"/",name,".txt",sep="")
+   file <- system.file(fileName, package = "DiffLogo")
+   motifs_dna[[name]] <- as.matrix(read.delim(file, FALSE))
+ }
> ## import DNA motifs for three transcription factors from table
> motif_folder <- "extdata/alignments"
> motif_names_dna2 <- c("Mad", "Max", "Myc")
> motifs_dna2 <- list()
> for (name in motif_names_dna2) {
+   fileName <- paste(motif_folder,"/",name,".txt",sep="")
+   file <- system.file(fileName, package = "DiffLogo")
+   fileContent <- readLines(file)
+   fileContent <- unlist(lapply(
+     X = fileContent,
```

```
+     FUN = function(x){ strsplit(x = x, split = "\t")[[1]][[1]] }))
+   motifs_dna2[[name]] <- getPwmFromAlignment(fileContent, DNA, 1)
+ }
> ## import three ASN motifs for one protein domain from fasta files
> motif_folder = "extdata/alignments"
> motif_names_asn = c("F-box_fungi.seq", "F-box_metazoa.seq",
+                     "F-box_viridiplantae.seq")
> motifs_asn = list()
> for (name in motif_names_asn) {
+   fileName = paste(motif_folder,"/",name,".fa",sep="")
+   file = system.file(fileName, package = "DiffLogo")
+   fileContent <- readLines(file)
+   fileContent <- fileContent[seq(from = 2, by = 2,
+                     length.out = floor(length(fileContent)/2))]
+   motifs_asn[[name]] <- getPwmFromAlignment(fileContent, ASN, 1)
+ }
```

Here, we import a set of nine DNA motifs for transcription factor CTCF from matrix files, a set of DNA motifs for three different E-Box transcription factors from sequences in tabular files, and a set of three ASN motifs for the F-Box protein domain from FASTA files.

4 Plot sequence logo

The user is able to examine motifs using the classical sequence logo from package *seqLogo* [7].

```
> ## plot classic sequence logo
> library(seqLogo)
> seqLogo::seqLogo(pwm = pwm1)
```

The user is also able to plot sequence logos with custom functions for stack height and base distribution using the package DiffLogo. In case of *stackHeight=informationContent* and *baseDistribution=probabilities*, the result is equivalent to the result of package *seqLogo*

```
> ## plot custom sequence logo
> par(mfrow=c(2,1), pin=c(3, 1), mar = c(2, 4, 1, 1))
```

```
> DiffLogo::seqLogo(pwm = pwm1)
> DiffLogo::seqLogo(pwm = pwm2, stackHeight = sumProbabilities)
> par(mfrow=c(1,1), pin=c(1, 1), mar=c(5.1, 4.1, 4.1, 2.1))
```

5 Plot difference logo

The user is easily able to plot a difference logo for a pair of motifs.

```
> ## plot DiffLogo
> diffLogoFromPwm(pwm1 = pwm1, pwm2 = pwm2)
> ## diffLogoFromPwm is a convenience function for
> diffLogoObj = createDiffLogoObject(pwm1 = pwm1, pwm2 = pwm2)
> diffLogo(diffLogoObj)
```

6 Plot table of difference logos

The user is easily able to plot a table of difference logos for a set of motifs.

```
> ## plot table of difference logos for CTFC motifs (DNA)
> diffLogoTable(PWMs = motifs_dna, )
> ## plot table of difference logos for E-Box motifs (DNA)
> diffLogoTable(PWMs = motifs_dna2)
> ## plot table of difference logos for F-Box motifs (ASN)
> diffLogoTable(PWMs = motifs_asn, alphabet = ASN)
```

7 Export visualization

The user is able to export the generated visualizations in various formats. Please find two examples below.

```
> ## parameters
> widthToHeightRatio = 16/10;
> size = length(motifs_dna) * 2
> resolution <- 300
> width <- size * widthToHeightRatio
> height <- size
> ## export single DiffLogo as pdf document
```

```
> fileName <- "Comparison_of_two_motifs.pdf"
> pdf(file = fileName, width = width, height = height)
> diffLogoFromPwm(pwm1 = pwm1, pwm2 = pwm2)
> dev.off()
```

```
pdf
  2
```

```
> ## export DiffLogo table as png image
> fileName <- "Comparison_of_multiple_motifs.png"
> png(
+   filename = fileName, res = resolution,
+   width = width * resolution, height = height * resolution)
> diffLogoTable(PWMs = motifs_dna)
> dev.off()
```

```
pdf
  2
```

Literature

[1] http://en.wikipedia.org/wiki/Position_weight_matrix
[2] Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18:6097-6100
[3] Shannon P (2014). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R package version 1.10.0.
[4] Eggeling, R., Gohr, A., Keilwagen, J., Mohr, M., Posch, S., Smith, A.D., Grosse, I.: On the value of intra-motifdependencies of human insulator protein ctcf. PLoS ONE 9(1), 85629 (2014). doi:10.1371/journal.pone.0085629
[5] Mordelet, Fantine and Horton, John and Hartemink, Alexander J and Engelhardt, Barbara E and Gordân, Raluca: Stability selection for regression-based models of transcription factor-DNA binding specificity. Bioinformatics 29(13), 11725 (2013). doi:10.1093/bioinformatics/btt221
[6] Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M.: Pfam: the protein families database. Nucleic Acids Research 42(D1), 222230 (2014). doi:10.1093/nar/gkt1223
[7] Bembom O. seqLogo: Sequence logos for DNA sequence alignments. R package version 1.34.0.