motifStack guide

Jianhong Ou[*], Lihua Julie Zhu[†]

March 6, 2013

Contents

1 Introduction **1**

2 Prepare environment **2**

3 Examples of using motifStack **2**
3.1 plot a DNA sequence logo with different fonts and colors . . . 2
3.2 plot an amino acid sequence logo 2
3.3 plot sequence logo stack 3

4 References **6**

5 Session Info **7**

1 Introduction

A sequence logo has been widely used as a graphical representation of an alignment of multiple amino acid or nucleic acid sequences. There is a package seqlogo[1] implemented in R to draw DNA sequence logos. However, it does not support amino acid sequence logos.

We have developed motifStack package for drawing sequence logos for protein, DNA and RNA sequences. motifStack provides the flexibility for users

[*]jianhong.ou@umassmed.edu
[†]Julie.Zhu@umassmed.edu

to select the font type and symbol colors. Comparing to seqlogo, motifStack has the capability for graphical representation of multiple motifs.

2 Prepare environment

You will need ghostscript: the full path to the executable can be set by the environment variable R_GSCMD. If this is unset, a GhostScript executable will be searched by name on your path. For example, on a Unix, linux or Mac "gs" is used for searching, and on Windows the setting of the environment variable GSC is used, otherwise commands "gswi64c.exe" then "gswin32c.exe" are tried.

Example on Windows: assume that the gswin32c.exe is installed at C:\Program Files\gs\gs9.06\bin, then open R and try: Sys.setenv(R_GSCMD="\"C:\\Program Files\\gs\\gs9.06\\bin\\gswin32c.exe\"")

3 Examples of using motifStack

3.1 plot a DNA sequence logo with different fonts and colors

Users can select different fonts and colors to draw the sequence logo.

```
> library(motifStack)
> pcm <- read.table(file.path(find.package("motifStack"), "extdata", "bin_SOLEXA.pcm"))
> pcm <- pcm[,3:ncol(pcm)]
> rownames(pcm) <- c("A","C","G","T")
> motif <- pcm2pfm(pcm)
> motif <- new("pfm", mat=motif, name="bin_SOLEXA")
> plot(motif)
> #try a different font
> plot(motif, font="mono,Courier")
> #try a different font and a different color group
> motif@color <- colorset(colorScheme='basepairing')
> plot(motif,font="Times")
```

3.2 plot an amino acid sequence logo

Given that motifStack allows to use any letters as symbols, it can also be used to draw amino acid sequence logos.

```
> library(motifStack)
> protein<-read.table(file.path(find.package("motifStack"),"extdata","cap.txt"))
```

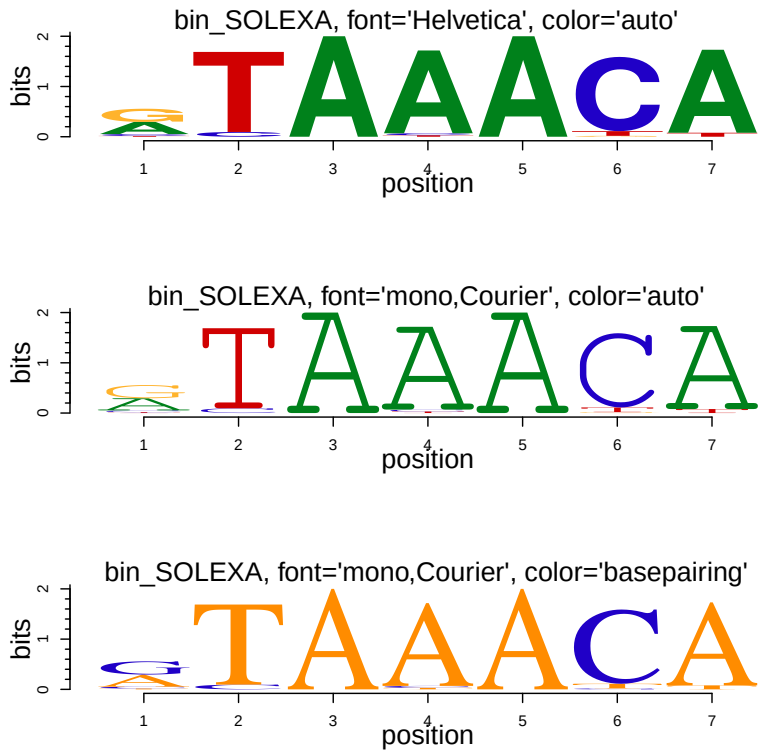

Figure 1: DNA sequence logo

```
> protein<-t(protein[,1:20])
> motif<-pcm2pfm(protein)
> motif<-new("pfm", mat=motif, name="CAP",
+             color=colorset(alphabet="AA",colorScheme="chemistry"))
> plot(motif)
```

3.3 plot sequence logo stack

motifStack is designed to show multiple motifs in same canvas. To show the sequence logo stack, motifs need to be aligned first for example by using MotIV[2]::motifMatch, which implemented STAMP[3]. After alignment, users can use plotMotifLogoStack function to draw sequence logos stack or use plotMotifLogoStackWithTree function to show the distance tree and sequence logos stack in the same canvas.

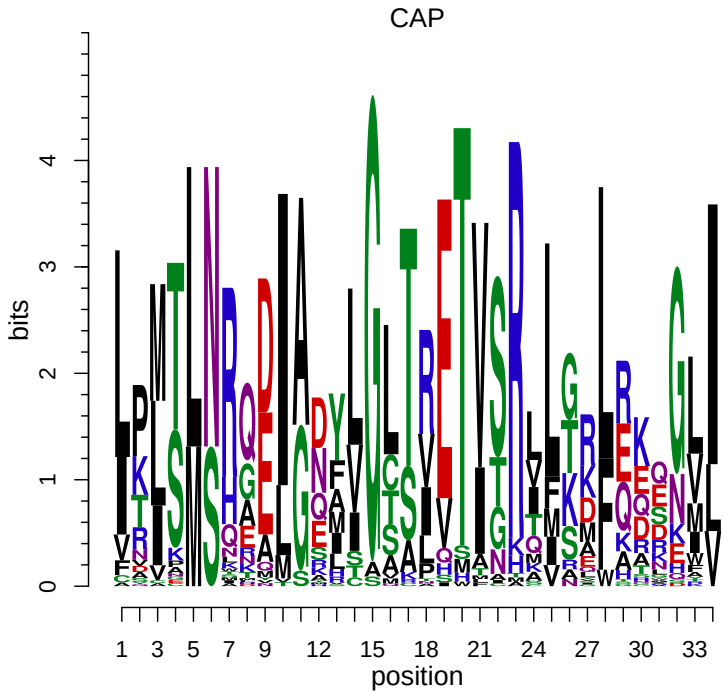

Figure 2: Amino acid sequence logo

```
> library(motifStack)
> library("MotIV")
> #####Database and Scores#####
> jaspar <- MotIV::readPWMfile(file.path(find.package("MotIV"), "extdata", "jaspar2010.txt"))
> jaspar.scores <- MotIV::readDBScores(file.path(find.package("MotIV"), "extdata", "jaspar2010_PCC_SWU.scores"))
> #####Input#####
> pcms<-readPCM(file.path(find.package("motifStack"), "extdata"),"pcm$")
> pcms<-lapply(pcms,function(.ele){.ele<-.ele[,3:ncol(.ele)];rownames(.ele)<-c("A","C","G","T");.ele})
> motifs<-lapply(pcms,pcm2pfm)
> #####Analysis#####
> foxa1.analysis.jaspar<-MotIV::motifMatch(inputPWM=motifs,
+                                          align="SWU",cc="PCC",
+                                          database=jaspar, DBscores=jaspar.scores,top=5)

        Ungapped Alignment
        Scores read
        Database read
        Motif matches : 5
```

```
> #####Clustering#####
> d <- MotIV::motifDistances(getPWM(foxa1.analysis.jaspar))
> hc <- MotIV::motifHclust(d)
> ##reorder the motifs for plotMotifLogoStack
> motifs<-motifs[hc$order]
> motifs<-lapply(names(motifs), function(.ele, motifs){new("pfm",mat=motifs[[.ele]], name=.ele)},motifs)
> ##do alignment
> motifs<-DNAmotifAlignment(motifs)
> ##plot stacks
> plotMotifLogoStack(motifs, ncex=1.0)
> plotMotifLogoStackWithTree(motifs, hc=hc)
```

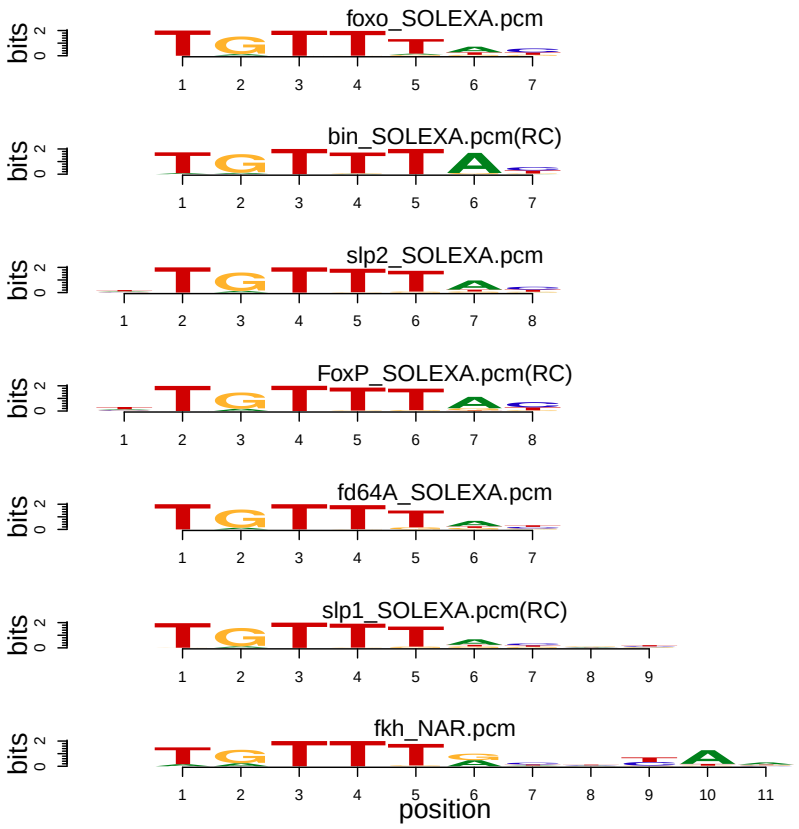

Figure 3: sequence logo stack

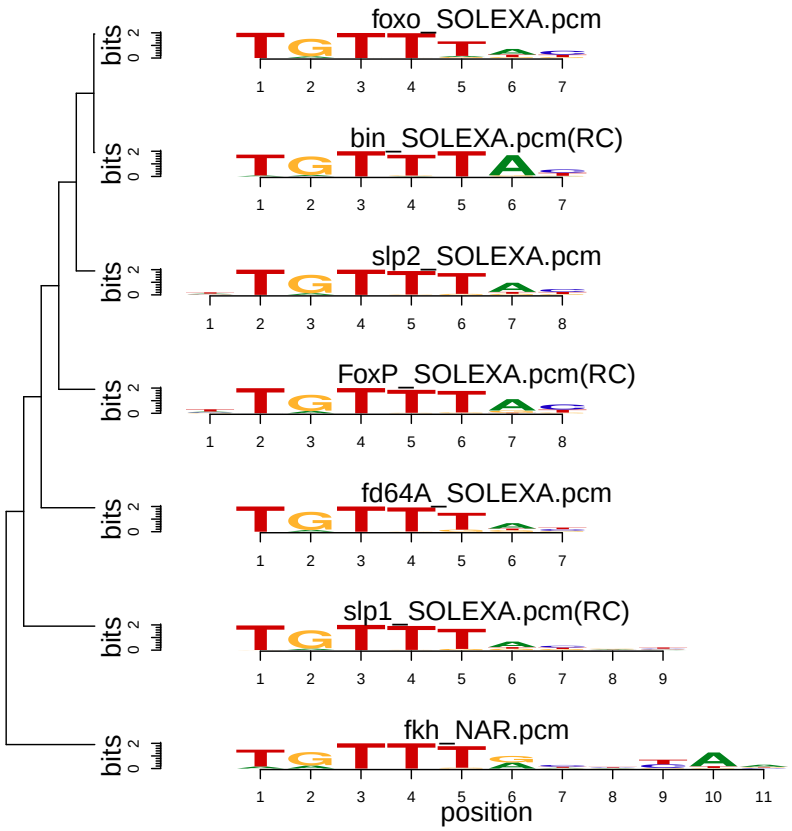

Figure 4: sequence logo stack with hierarchical cluster tree

4 References

References

[1] seqLogo: Sequence logos for DNA sequence alignments. R package version 1.22.0.

[2] MotIV: Motif Identification and Validation. Eloi Mercier and Raphael Gottardo (2010). R package version 1.10.0.

[3] STAMP: a web tool for exploring DNA-binding motif similarities. Mahony S, Benos PV, Nucleic Acids Res. 2007, 35(Web Server issue): W253-W258.

5 Session Info

```
> sessionInfo()

R version 2.15.3 (2013-03-01)
Platform: i386-w64-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] MotIV_1.14.0       BiocGenerics_0.4.0 motifStack_1.0.4   grImport_0.8-4
[5] XML_3.95-0.1

loaded via a namespace (and not attached):
 [1] BSgenome_1.26.1      Biostrings_2.26.3    GenomicRanges_1.10.7
 [4] IRanges_1.16.6       lattice_0.20-13      parallel_2.15.3
 [7] rGADEM_2.6.0         seqLogo_1.24.0       stats4_2.15.3
[10] tools_2.15.3
```