motifStack guide

Jianhong Ou[*], Lihua Julie Zhu[†]

April 10, 2013

Contents

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.

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

We have developed motifStack package for drawing sequence logos for protein, DNA and RNA sequences. motifStack provides the flexibility for users 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 <- new("pcm", mat=as.matrix(pcm), name="bin_SOLEXA")
> ##pfm object
> #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.

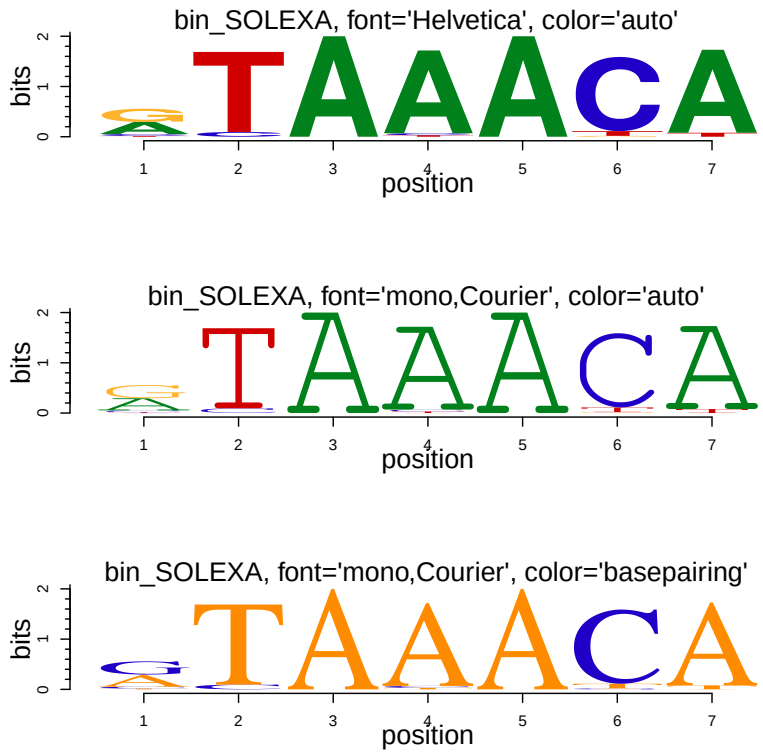

Figure 1: DNA sequence logo

```
> library(motifStack)
> protein<-read.table(file.path(find.package("motifStack"),"extdata","cap.txt"))
> 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, the distance of motifs need to be calculated first for example by using MotIV[2]::motifDistances, which implemented STAMP[3]. After alignment, users can use plotMotifLogoStack function to draw sequence logos stack or use plotMotifLogoStackWithTree function to show the distance

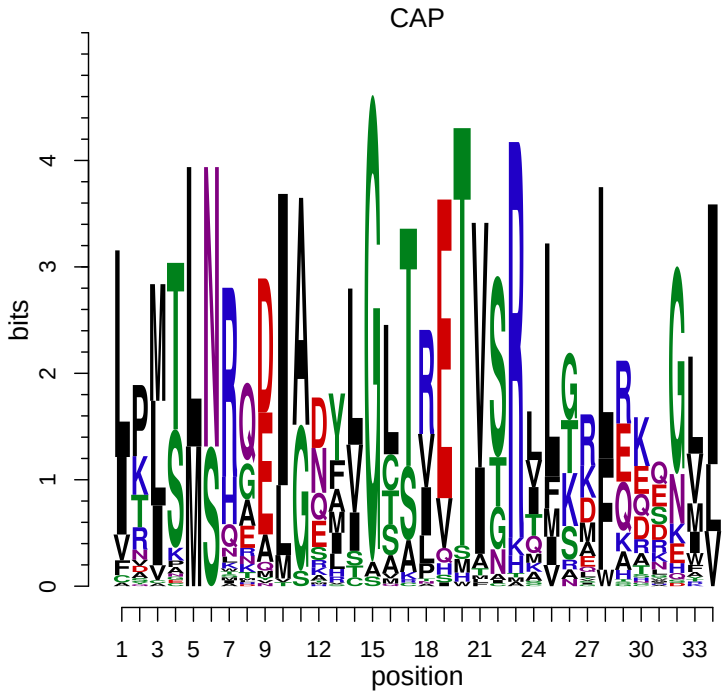

Figure 2: Amino acid sequence logo

tree with the sequence logos stack or use plotMotifStackWithRadialPhylog function to plot sequence logo stack in radial style in the same canvas. There is a shortcut function named as motifStack. Use stack layout to call plotMotifLogoStack, treeview layout to call plotMotifLogoStackWithTree and radialPhylog to call plotMotifStackWithRadialPhylog.

```
> library(motifStack)
> #####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)
> motifs<-lapply(names(motifs), function(.ele, motifs){new("pfm",mat=motifs[[.ele]], name=.ele)},motifs)
> ##plot stacks
> motifStack(motifs, layout="stack", ncex=1.0)
> motifStack(motifs, layout="tree")
> ###When the number of motifs is too much to be shown in a vertical stack,
> ###motifStack can draw them in a radial style.
```

```
> library("MotifDb")
> matrix.fly <- query(MotifDb, "Dmelanogaster")
> motifs2 <- as.list(matrix.fly)
> motifs2 <- motifs2[grepl("Dmelanogaster\\-FlyFactorSurvey\\-", names(motifs2))]
> names(motifs2) <- gsub("Dmelanogaster_FlyFactorSurvey_", "",
+                        gsub("_FBgn\\d+$", "",
+                             gsub("[^a-zA-Z0-9]","_",
+                                  gsub("(_\\d+)+$", "", names(motifs2)))))
> motifs2 <- motifs2[unique(names(motifs2))]
> pfms <- sample(motifs2, 50)
> motifs2 <- lapply(names(pfms), function(.ele, pfms){new("pfm",mat=pfms[[.ele]], name=.ele)},pfms)
> library(RColorBrewer)
> color <- brewer.pal(12, "Set3")

>motifStack(motifs2, layout="radialPhylog",
+                                col.bg=rep(color, each=5), col.bg.alpha=0.3,
+                                col.leaves=rep(color, each=5),
+                                col.inner.label.circle=rep(color,5),
+                                col.outer.label.circle=rep(color,5), outer.label.circle.width=0.1,
+                                angle=350)
```

3.4 plot a sequence logo cloud

We can also plot a sequence logo cloud for DNA sequence logo.

```
> groups <- rep(paste("group",1:5,sep=""), each=10)
> names(groups) <- names(pfms)
> group.col <- brewer.pal(5, "Set3")
> names(group.col)<-paste("group",1:5,sep="")
> jaspar.scores <- MotIV::readDBScores(file.path(find.package("MotIV"), "extdata", "jaspar2010_PCC_SWU.scores"))
> d <- MotIV::motifDistances(pfms)
> hc <- MotIV::motifHclust(d)
> phylog <- hclust2phylog(hc)
> leaves <- names(phylog$leaves)
> pfms <- pfms[leaves]
> pfms <- lapply(names(pfms), function(.ele, pfms){new("pfm",mat=pfms[[.ele]], name=.ele)},pfms)
> motifSig <- motifSignature(pfms, phylog, groupDistance=0.1)

> motifCloud(motifSig, scale=c(6, .5),
+                        layout="rectangles", group.col=group.col, groups=groups, draw.legend=T)
```

4 References

References

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

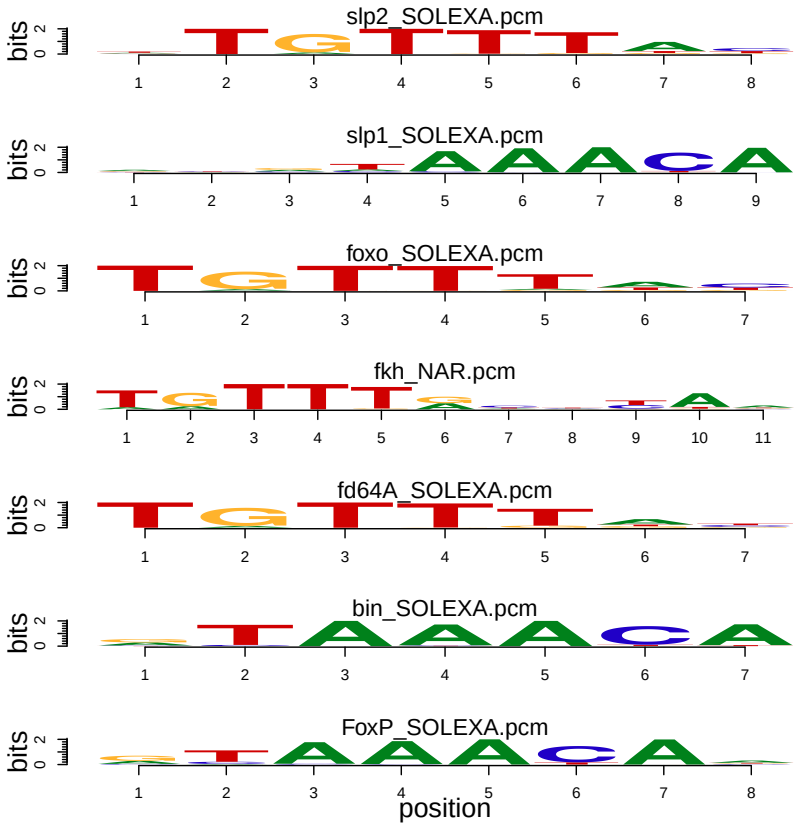

Figure 3: sequence logo stack

[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 3.0.0 (2013-04-03)
Platform: x86_64-apple-darwin10.8.0 (64-bit)
```

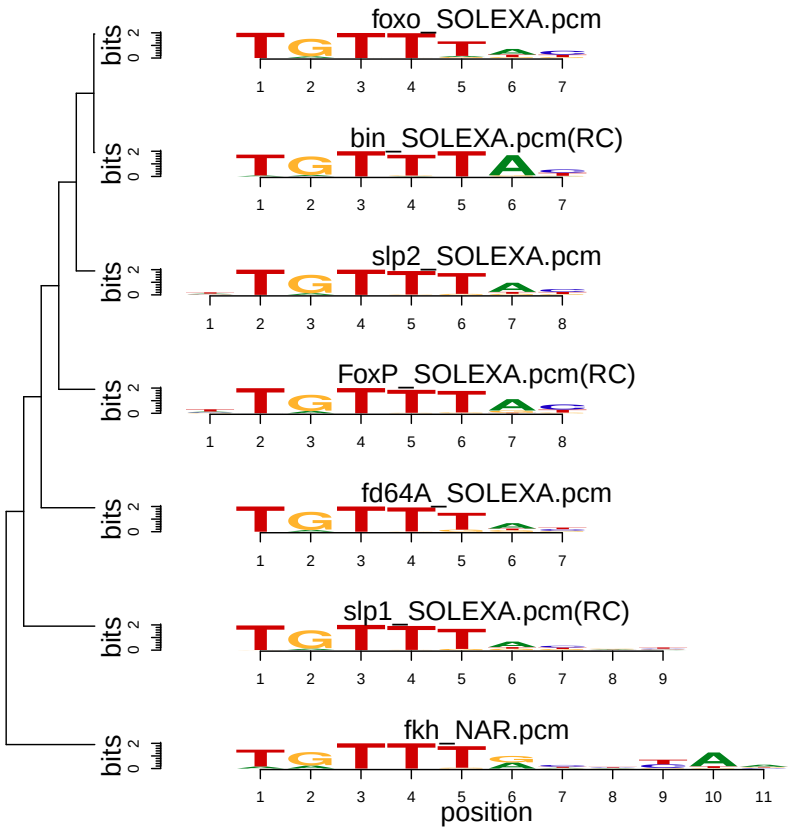

Figure 4: sequence logo stack with hierarchical cluster tree

```
locale:
[1] C

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

other attached packages:
 [1] RColorBrewer_1.0-5 MotifDb_1.2.0      Biostrings_2.28.0  IRanges_1.18.0
 [5] motifStack_1.4.0   ade4_1.5-1         MotIV_1.16.0       BiocGenerics_0.6.0
 [9] grImport_0.8-4     XML_3.96-1.1
```

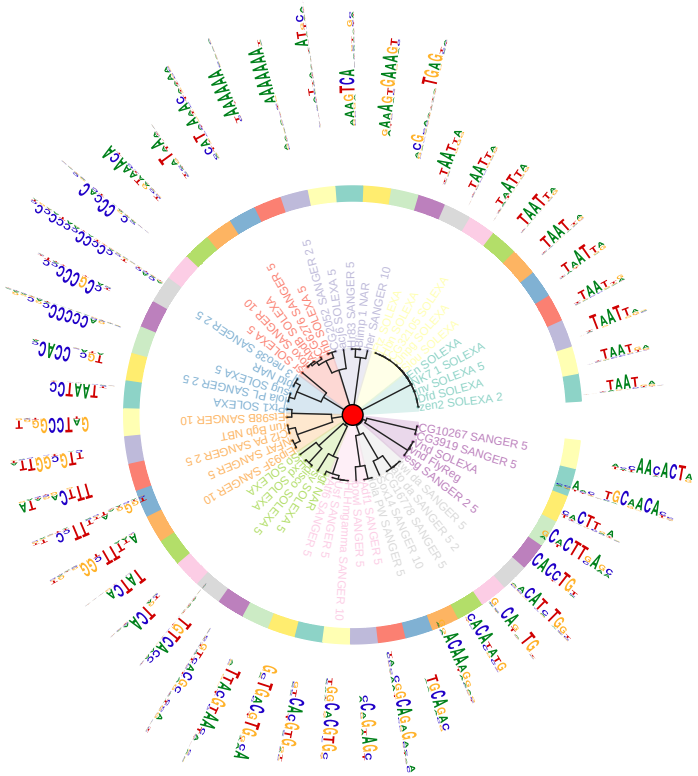

Figure 5: sequence logo stack in radial style

```
loaded via a namespace (and not attached):
 [1] BSgenome_1.28.0       GenomicRanges_1.12.1 RCurl_1.95-4.1
 [4] Rsamtools_1.12.0      bitops_1.0-5         lattice_0.20-15
 [7] rGADEM_2.8.0          rtracklayer_1.20.0   seqLogo_1.26.0
[10] stats4_3.0.0          tools_3.0.0          zlibbioc_1.6.0
```

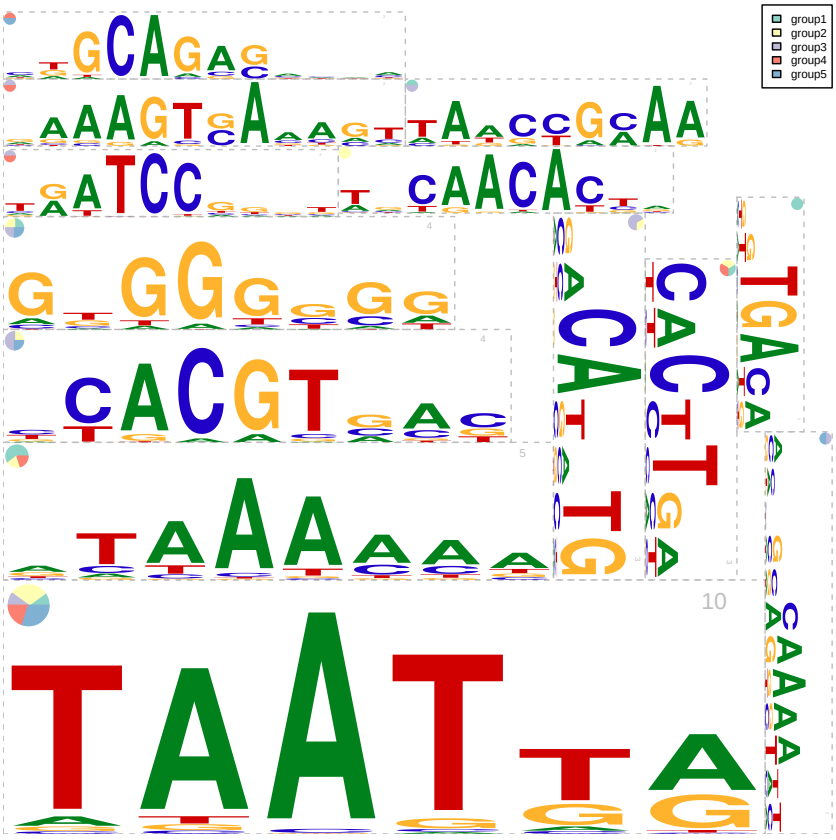

Figure 6: a sequence logo cloud with rectangle packing layout