motifStack guide

Jianhong Ou, Lihua Julie Zhu

October 13, 2015

Contents

1 Introduction

A sequence logo, based on information theory, has been widely used as a graphical representation of sequence conservation (aka motif) in multiple amino acid or nucleic acid sequences. Sequence motif represents conserved characteristics such as DNA binding sites, where transcription factors bind, and catalytic sites in enzymes. Although many tools, such as seqlogo[1], have been developed to create sequence motif and to represent it as individual sequence logo, software tools for depicting the relationship among multiple sequence motifs are still lacking. We developed a flexible and powerful open-source R/Bioconductor package, motifStack, for visualization of the alignment of multiple sequence 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=file.path("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 (Figure 1).

```
suppressPackageStartupMessages(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")
opar<-par(mfrow=c(4,1))
plot(motif)
#plot the logo with same height
plot(motif, ic.scale=FALSE, ylab="probability")
#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")
par(opar)
```

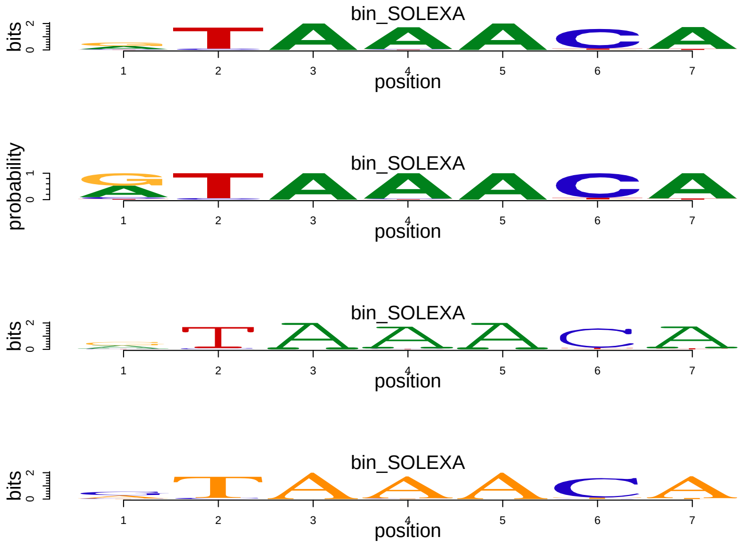

Figure 1: **DNA sequence logo.** Plot a DNA sequence logo with different fonts and colors.

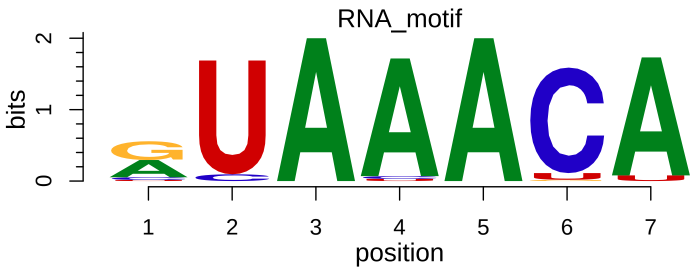

Figure 2: **RNA sequence logo.** Plot an RNA sequence logo

3.2 plot a RNA sequence logo

From DNA sequence logo to RNA sequence logo (Figure 2), you just need to change the rowname of the matrix from "T" to "U".

```
rna <- pcm
rownames(rna)[4] <- "U"
motif <- new("pcm", mat=as.matrix(rna), name="RNA_motif")
plot(motif)
```

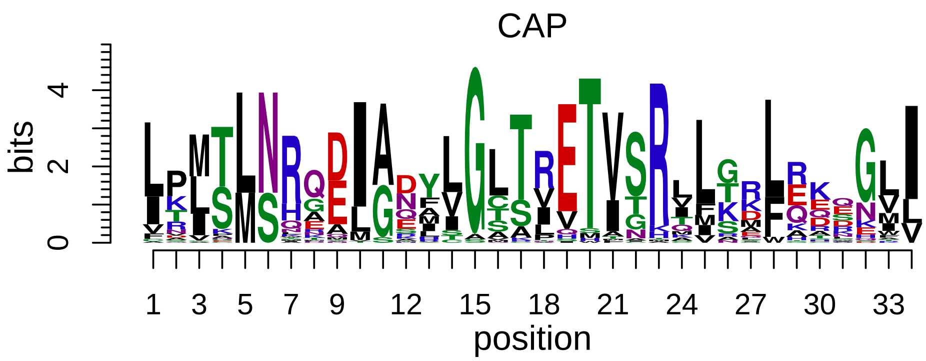

Figure 3: **Amino acid sequence logo.** Plot an sequence logo with any symbols as you want such as amino acid sequence logo

3.3 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 3).

```
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)
```

3.4 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 (Figure 4) or use plotMotifLogoStackWithTree function to show the distance tree with the sequence logos stack (Figure 5) or use plotMotifStackWithRadialPhylog function to plot sequence logo stack in radial style (Figure 6) 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#####
```

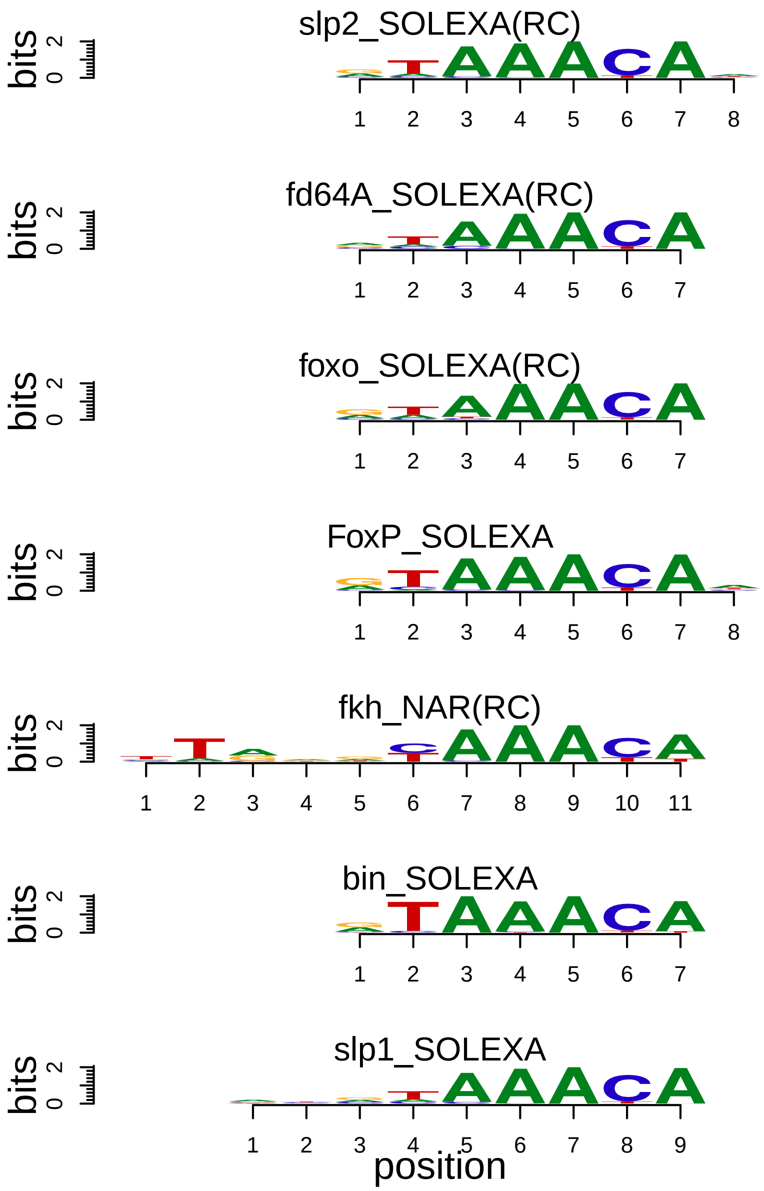

Figure 4: **Sequence logo stack.** Plot motifs with sequence logo stack style.

```
pcms<-readPCM(file.path(find.package("motifStack"), "extdata"),"pcm$")
motifs<-lapply(pcms,pcm2pfm)

## plot stacks
motifStack(motifs, layout="stack", ncex=1.0)
```

```
## plot stacks with hierarchical tree
motifStack(motifs, layout="tree")
```

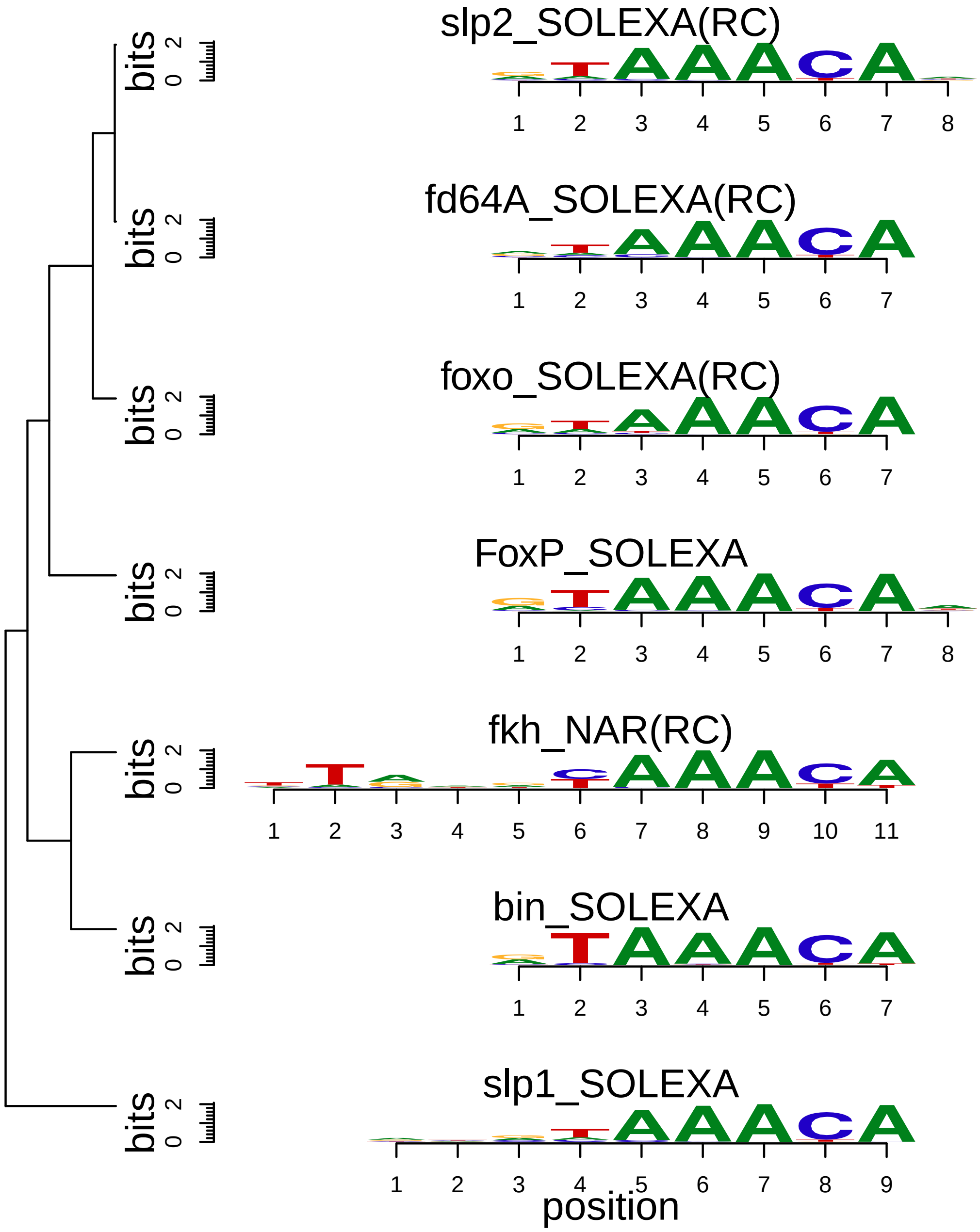

Figure 5: **Treeview layout logo stack.** Sequence logo stack with hierarchical cluster tree.

```
## When the number of motifs is too much to be shown in a vertical stack,
## motifStack can draw them in a radial style.
## random sample from MotifDb
library("MotifDb")
matrix.fly <- query(MotifDb, "Dmelanogaster")
motifs2 <- as.list(matrix.fly)
## use data from FlyFactorSurvey
motifs2 <- motifs2[grepl("Dmelanogaster\\-FlyFactorSurvey\\-",
                         names(motifs2))]
## format the names
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)
## creat a list of object of pfm
motifs2 <- lapply(names(pfms),
                  function(.ele, pfms){new("pfm",mat=pfms[[.ele]], name=.ele)}
                  ,pfms)
## trim the motifs
motifs2 <- lapply(motifs2, trimMotif, t=0.4)
## setting colors
library(RColorBrewer)
color <- brewer.pal(12, "Set3")
## plot logo stack with radial style
motifStack(motifs2, layout="radialPhylog",
           circle=0.3, cleaves = 0.2,
           clabel.leaves = 0.5,
           col.bg=rep(color, each=5), col.bg.alpha=0.3,
           col.leaves=rep(color, each=5),
           col.inner.label.circle=rep(color, each=5),
           inner.label.circle.width=0.05,
           col.outer.label.circle=rep(color, each=5),
           outer.label.circle.width=0.02,
           circle.motif=1.2,
           angle=350)
```

3.5 plot a sequence logo cloud

We can also plot a sequence logo cloud for DNA sequence logo (Figure 7).

```
## assign groups for motifs
groups <- rep(paste("group",1:5,sep=""), each=10)
names(groups) <- names(pfms)
## assign group colors
group.col <- brewer.pal(5, "Set3")
names(group.col)<-paste("group",1:5,sep="")
## use MotIV to calculate the distances of motifs
jaspar.scores <- MotIV::readDBScores(file.path(find.package("MotIV"),
                                               "extdata",
```

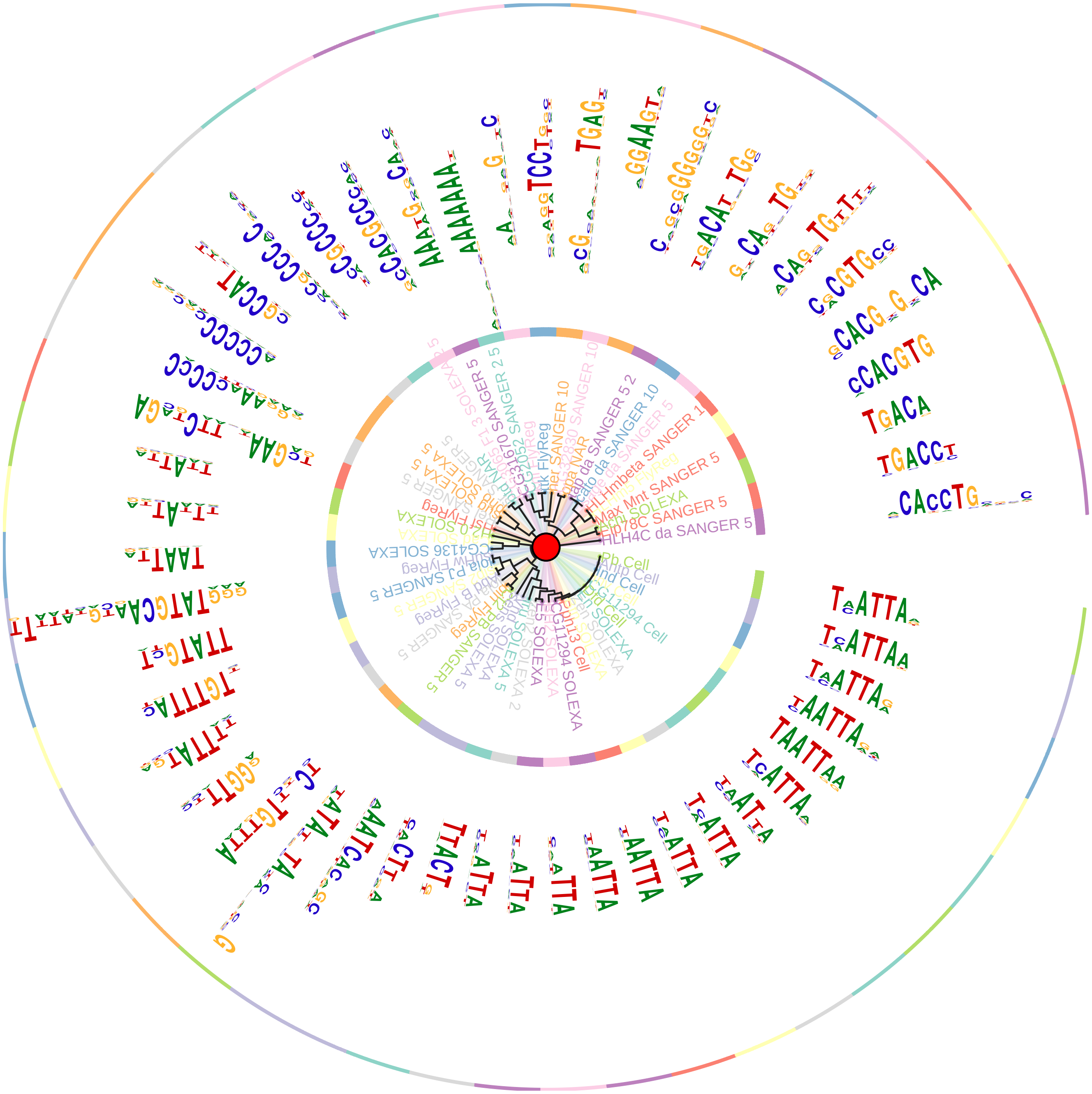

Figure 6: **Sequence logo stack in radial style** Plot motifs in a radial style when the number of motifs is too much to be shown in a vertical stack.

```
                                              "jaspar2010_PCC_SWU.scores"))
d <- MotIV::motifDistances(lapply(pfms, pfm2pwm))
hc <- MotIV::motifHclust(d, method="average")
## convert the hclust to phylog object
phylog <- hclust2phylog(hc)
## reorder the pfms by the order of hclust
```

```
leaves <- names(phylog$leaves)
pfms <- pfms[leaves]
## create a list of pfm objects
pfms <- lapply(names(pfms), function(.ele, pfms){
                                new("pfm",mat=pfms[[.ele]], name=.ele)}
               ,pfms)
## extract the motif signatures
motifSig <- motifSignature(pfms, phylog, groupDistance=0.01, min.freq=1)
## draw the motifs with a tag-cloud style.
motifCloud(motifSig, scale=c(6, .5),
           layout="rectangles",
           group.col=group.col,
           groups=groups,
           draw.legend=TRUE)
```

3.6 plot grouped sequence logo

To plot grouped sequence logo, except do motifCloud, we can also plot it with radialPhylog style (Figure 8).

```
## get the signatures from object of motifSignature
sig <- signatures(motifSig)
## set the inner-circle color for each signature
gpCol <- sigColor(motifSig)
## plot the logo stack with radial style.
plotMotifStackWithRadialPhylog(phylog=phylog, pfms=sig,
                              circle=0.4, cleaves = 0.3,
                              clabel.leaves = 0.5,
                              col.bg=rep(color, each=5), col.bg.alpha=0.3,
                              col.leaves=rep(rev(color), each=5),
                              col.inner.label.circle=gpCol,
                              inner.label.circle.width=0.03,
                              angle=350, circle.motif=1.2,
                              motifScale="logarithmic")
```

3.7 motifCircos

We can also plot it with circos style (Figure 9). In circos style, we can plot two group of motifs and with multiple color rings.

```
## plot the logo stack with radial style.
motifCircos(phylog=phylog, pfms=pfms, pfms2=sig,
            col.tree.bg=rep(color, each=5), col.tree.bg.alpha=0.3,
```

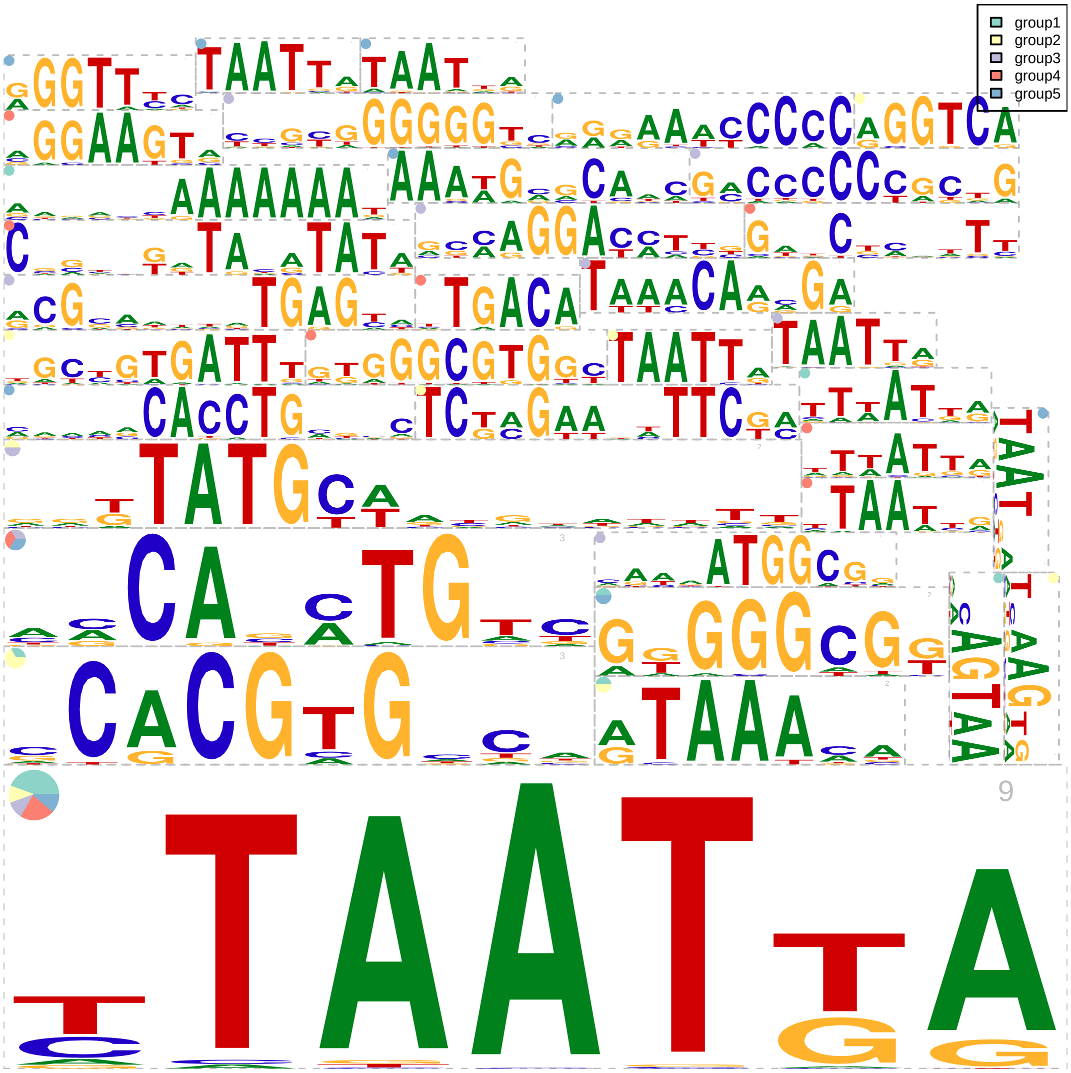

Figure 7: **Sequence logo cloud with rectangle packing layout** Like tag-cloud, the sequence logo size is determined by the number of motifs of the signature. The group sources of the motifs for each signature are shown as a pie graph in topleft corner.

```
            col.leaves=rep(rev(color), each=5),
            col.inner.label.circle=gpCol,
            inner.label.circle.width=0.03,
            col.outer.label.circle=gpCol,
            outer.label.circle.width=0.03,
            r.rings=c(0.02, 0.03, 0.04),
            col.rings=list(sample(colors(), 50),
                           sample(colors(), 50),
                           sample(colors(), 50)),
            angle=350, motifScale="logarithmic")
```

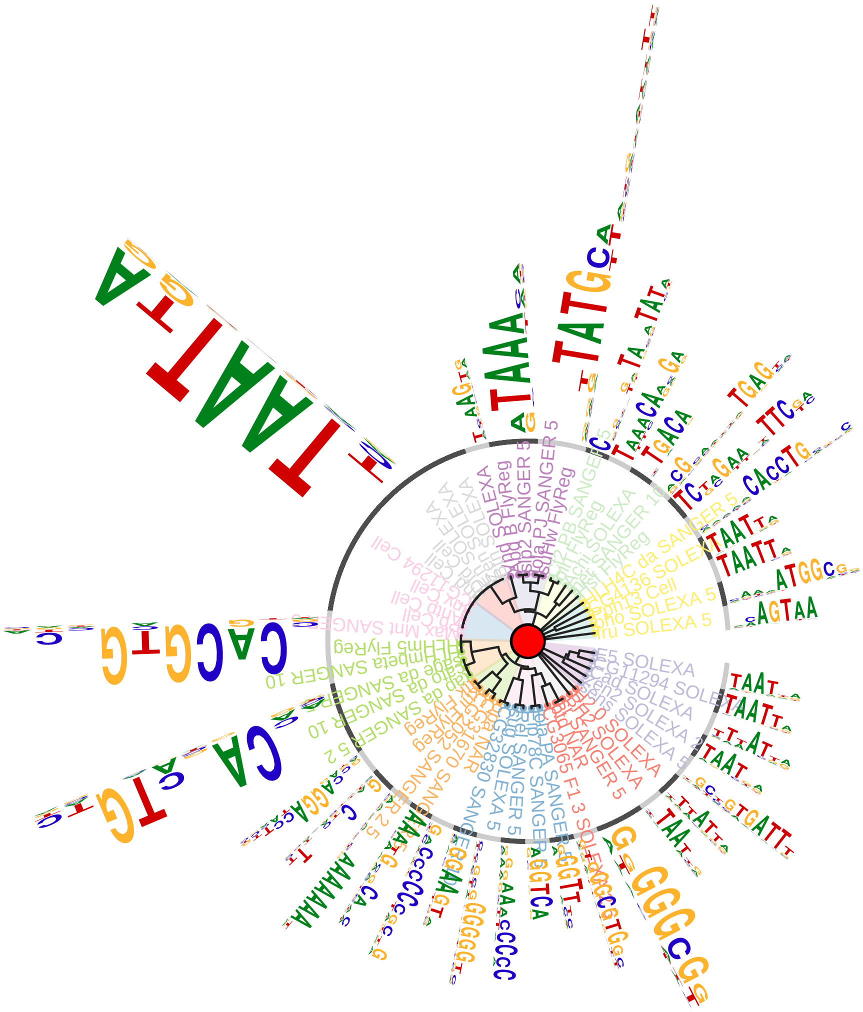

Figure 8: **Grouped sequence logo with radialPhylog style layout.** Like tag-cloud, the sequence logo size is determined by the number of motifs for the signature. The gray-black circle indicates the range of each signature.

3.8 motifPiles

We can also plot it with pile style (Figure 10). In pile style, we can plot two group of motifs and with multiple color annoations.

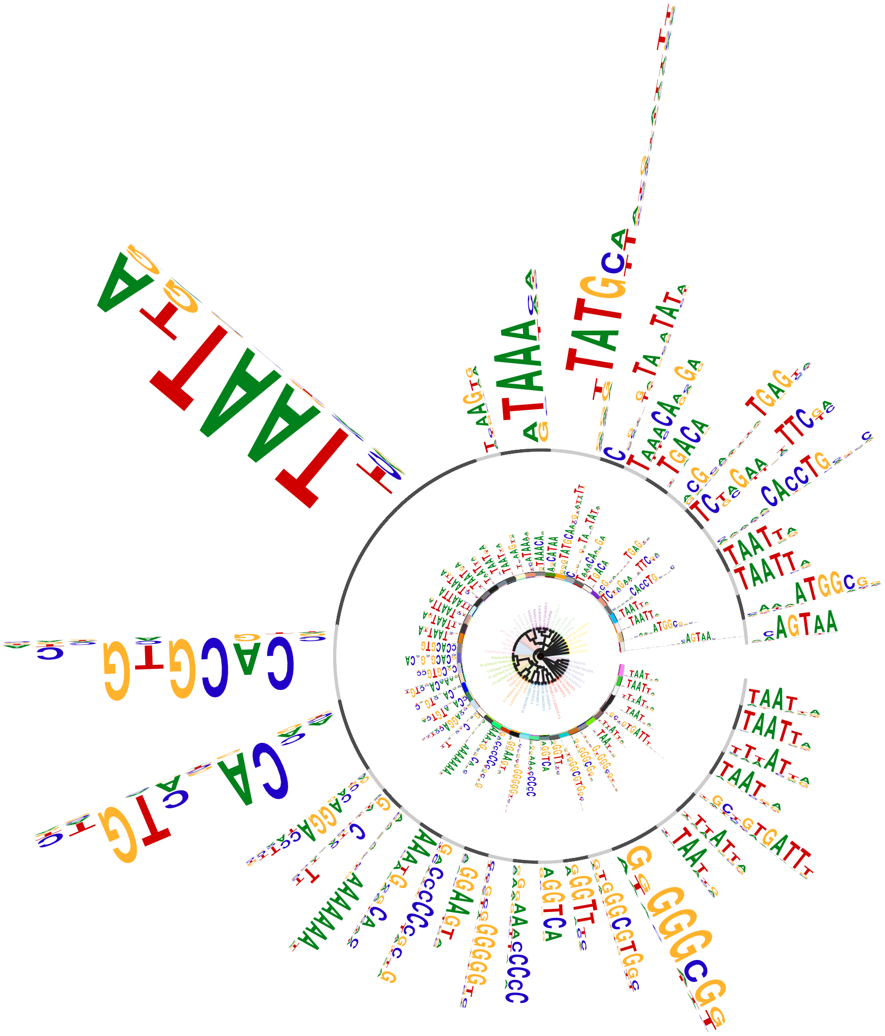

Figure 9: **Grouped sequence logo with circos style layout.** more color sets with more motifs.

```
## plot the logo stack with radial style.
motifPiles(phylog=phylog, pfms=pfms, pfms2=sig,
            col.tree=rep(color, each=5),
            col.leaves=rep(rev(color), each=5),
            col.pfms2=gpCol,
            r.anno=c(0.02, 0.03, 0.04),
```

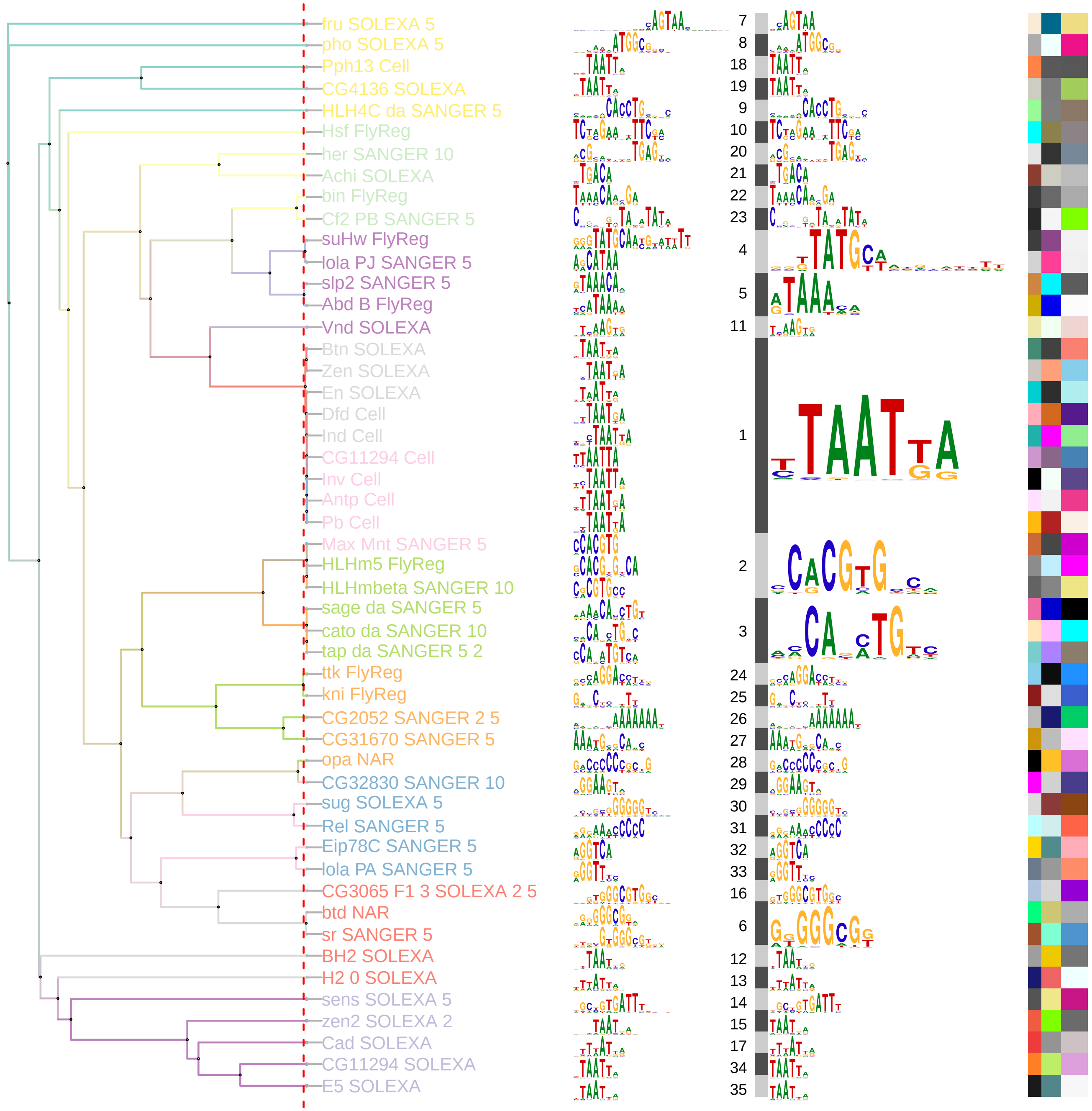

Figure 10: **Grouped sequence logo with piles style layout.** more color sets with more motifs.

```
             col.anno=list(sample(colors(), 50),
                           sample(colors(), 50),
                           sample(colors(), 50)),
             motifScale="logarithmic",
             plotIndex=TRUE,
             groupDistance=0.01)
```

4 docker container for motifStack

Docker allows software to be packaged into containers and the containers can be run any platform as well using a virtual machine called boot2docker. motifStack has its docker image stored in Docker Hub. Users can download the image and run.

```
docker pull jianhong/motifstack_1.13.6
cd ˜ ## in windows, please try cd c:\textbackslash Users\textbackslash username
mkdir tmp4motifstack ## this will be the share folder for your host and container.
docker run -ti --rm -v ${PWD}/tmp4motifstack:/volume/data jianhong/motifstack_1.13.6 R
## in R
setwd("/tmp")
library(motifStack)
packageVersion("motifStack")
pcmpath <- "pcmsDatasetFly"
pcms <- readPCM(pcmpath)
pfms <- lapply(pcms, pcm2pfm)
matalign_path <- "/usr/bin/matalign"
neighbor_path <- "/usr/bin/phylip/neighbor"
outpath <- "output"
system(paste("perl MatAlign2tree.pl --in . --pcmpath", pcmpath, "--out", outpath,
    "--matalign", matalign_path, "--neighbor", neighbor_path, "--tree","UPGMA"))
newickstrUPGMA <- readLines(con=file.path(outpath, "NJ.matalign.distMX.nwk"))
phylog <- newick2phylog(newickstrUPGMA, FALSE)
leaves <- names(phylog$leaves)
motifs <- pfms[leaves]
motifSig <- motifSignature(motifs, phylog, groupDistance=2, min.freq=1, trim=.2)
sig <- signatures(motifSig)
gpCol <- sigColor(motifSig)
leaveNames <- gsub("^Dm_", "", leaves)
pdf("/volume/data/test.pdf", width=8, height=11)
motifPiles(phylog=phylog, DNAmotifAlignment(motifs), sig,
    col.pfms=gpCol, col.pfms.width=.01,
    col.pfms2=gpCol, col.pfms2.width=.01,
    labels.leaves=leaveNames,
    plotIndex=c(FALSE, TRUE), IndexCex=1,
    groupDistance=2, clabel.leaves=1)
dev.off()
```

You will see the test.pdf file in the folder of tmp4motifstack.

5 References

References

[1] Oliver Bembom. seqlogo: Sequence logos for dna sequence alignments. *R package version 1.5.4*, 2006.

[2] Eloi Mercier and Raphael Gottardo. Motiv: Motif identification and validation. *R package version 1.10.0*, 2010.

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

6 Session Info

```
sessionInfo()

## R version 3.2.2 (2015-08-14)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.3 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
##  [4] LC_COLLATE=C               LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
##  [1] stats4    parallel  grid      stats     graphics  grDevices utils     datasets
##  [9] methods   base
##
## other attached packages:
##  [1] RColorBrewer_1.1-2  MotifDb_1.12.0      motifStack_1.14.0   Biostrings_2.38.0
##  [5] XVector_0.10.0      IRanges_2.4.0       S4Vectors_0.8.0     ade4_1.7-2
##  [9] MotIV_1.26.0        BiocGenerics_0.16.0 grImport_0.9-0      XML_3.98-1.3
## [13] BiocStyle_1.8.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.1                highr_0.5.1                plyr_1.8.3
##  [4] formatR_1.2.1              futile.logger_1.4.1        GenomeInfoDb_1.6.0
##  [7] bitops_1.0-6               futile.options_1.0.0       tools_3.2.2
## [10] zlibbioc_1.16.0            digest_0.6.8               evaluate_0.8
## [13] lattice_0.20-33            BSgenome_1.38.0            yaml_2.1.13
## [16] seqLogo_1.36.0             rtracklayer_1.30.0         stringr_1.0.0
## [19] knitr_1.11                 Biobase_2.30.0             BiocParallel_1.4.0
## [22] rGADEM_2.18.0              rmarkdown_0.8.1            lambda.r_1.1.7
## [25] magrittr_1.5                Rsamtools_1.22.0            scales_0.3.0
## [28] htmltools_0.2.6             GenomicRanges_1.22.0        GenomicAlignments_1.6.0
## [31] SummarizedExperiment_1.0.0 colorspace_1.2-6            stringi_0.5-5
## [34] munsell_0.4.2               RCurl_1.95-4.7
```