

motifStack 简介

欧剑虹*, Lihua Julie Zhu†

2012 年 10 月 16 日

目录

1 背景	1
2 准备工作	2
3 快速上手	2
3.1 使用不同的字体及颜色绘制序列图标	2
3.2 绘制氨基酸序列图标	2
3.3 绘制图标堆	3
3.4 绘制环形风格的图标堆	5
4 References	7
5 Session Info	8

1 背景

序列图标是用于图形化显示氨基酸或者核酸模序的手段. 在 Bioconductor 中 seqlogo[1] 是生成序列图标的一个包, 但是它不能用于氨基酸模序的显示.

motifStack 包是为了方便人们在 R 中画出蛋白蛋,DNA, 以及 RNA 序列图标. motifStack 可以让人们自由地选择自己喜欢的字体和颜色来绘制图

*jianhong.ou@umassmed.edu

†Julie.Zhu@umassmed.edu

标. 与 seqlogo 不同的是,motifStack 还可以同时在同一画布上绘制多个图标.

2 准备工作

motifStack 需要 GhostScript 才能正常运行. 在安装了 GhostScript 之后, 你需要设置 R 中的 R_GSCMD 环境变量. 通常安装好的 R 会自动设置好这个变量. 如果未能正常设置, 请参考下面的例子设置即可.

比如你使用的是 Windows 32 位操作系统,GhostScript 安装在 C:\Program Files\gs\gs9.06\bin, 那么你打开 R 后输入: `Sys.setenv(R_GSCMD="\"C:\\Program Files\\gs\\gs9.06\\bin\\gswin32c.exe\"")`

3 快速上手

3.1 使用不同的字体及颜色绘制序列图标

用户可以自由地选择自己喜欢的字体和颜色绘制图标.

```
\begin{lstlisting}[basicstyle=\ttfamily]
> 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)
> # 实例化 pfm 类
> motif <- new("pfm", mat=motif, name="bin_SOLEXA")
> plot(motif)
> # 尝试使用不同的字体
> plot(motif, font="mono,Courier")
> #t 尝试使用不同的颜色
> motif@color <- colorset(colorScheme='basepairing')
> plot(motif,font="Times")
\end{lstlisting}
```

3.2 绘制氨基酸序列图标

实际上,motifStack 允许用户使用任何字符做为图标中的字母, 这是由 pfm 类实例的 mat 行号决定. 所以它可以被用来绘制氨基酸序列图标.

```
> 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",
```

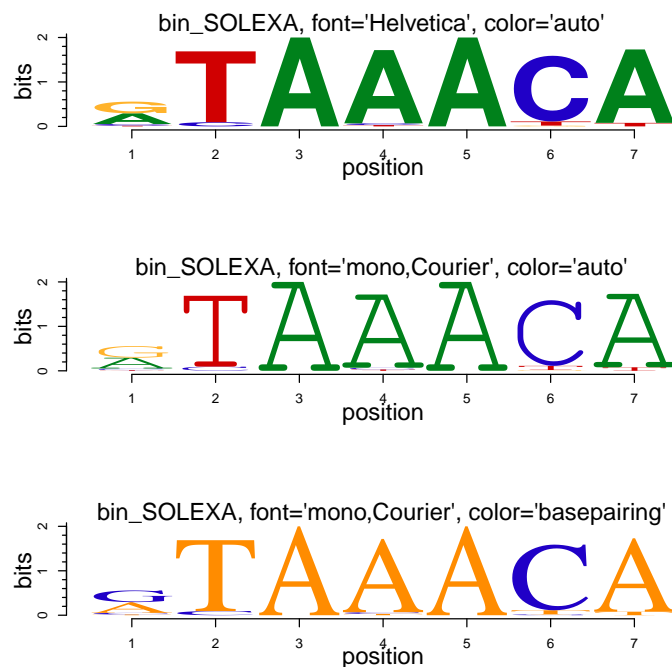


图 1: DNA 序列图标

```
+ color=colorset(alphabet="AA",colorScheme="chemistry")
> plot(motif)
```

3.3 绘制图标堆

motifStack 可以在同一画布上绘制多个图标. 如果图标需要对齐绘制的话, 需要事先对模序进行对齐工作. 这一步可以由整合了 STAMP[3] 的 MotIV[2]::motifMatch 来完成. 对齐以后就可以使用 plotMotifLogoStack() 以及 plotMotifLogoStackWithTree() 函数来绘制图标堆以及左侧带有进化树的图标堆了.

```
> library(motifStack)
> library("MotIV")
> ##### 读入数据库及打分表 #####
> 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"))
> ##### 输入 #####
```

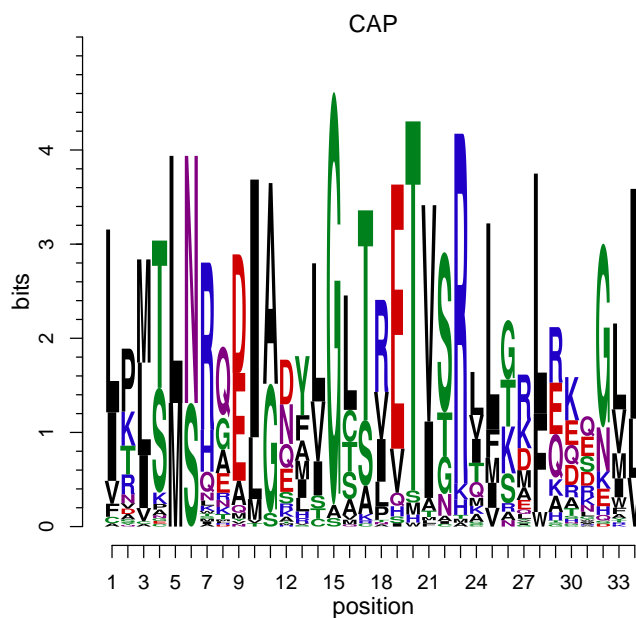


图 2: 氨基酸序列图标

```

> 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)
> ##### 分析 #####
> foxal.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

> ##### 聚类分析 #####
> d <- MotIV::motifDistances(getPWM(foxal.analysis.jaspar))
> hc <- MotIV::motifHclust(d)
> ## 对模式进行重排
> motifs<-motifs[hc$order]
> motifs<-lapply(names(motifs), function(.ele, motifs){new("pfm",mat=motifs[[.ele]], name=.ele)},motifs)
> ## 对齐
> motifs<-DNAmotifAlignment(motifs)
> ## 绘制图形

```

```
> plotMotifLogoStack(motifs, ncex=1.0)
> plotMotifLogoStackWithTree(motifs, hc=hc)
```

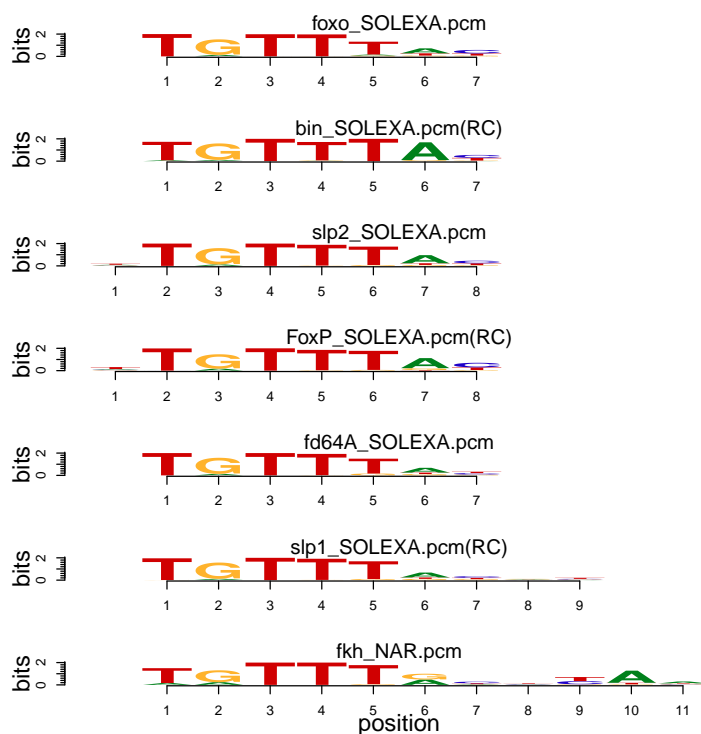


图 3: 序列图标堆

3.4 绘制环形风格的图标堆

当图标过多, 无法使用上面的函数绘制的时候, 可以使用 `plotMotifStackWithRadialPhylog()` 函数绘制环形风格图标堆.

```
> library("MotifDb")
> library("MotIV")
> library("motifStack")
> matrix.fly <- query(MotifDb, "Dmelanogaster")
> motifs <- as.list(matrix.fly)
> motifs <- motifs[grepl("Dmelanogaster\\-FlyFactorSurvey\\-", names(motifs))]
> names(motifs) <- gsub("Dmelanogaster_FlyFactorSurvey_", "",
+                       gsub("_FBgn\\d+$", "",
+                             gsub("[^a-zA-Z0-9]", "_",
+                                   gsub("(\\d+)$", "", names(motifs))))))
> motifs <- motifs[unique(names(motifs))]
```

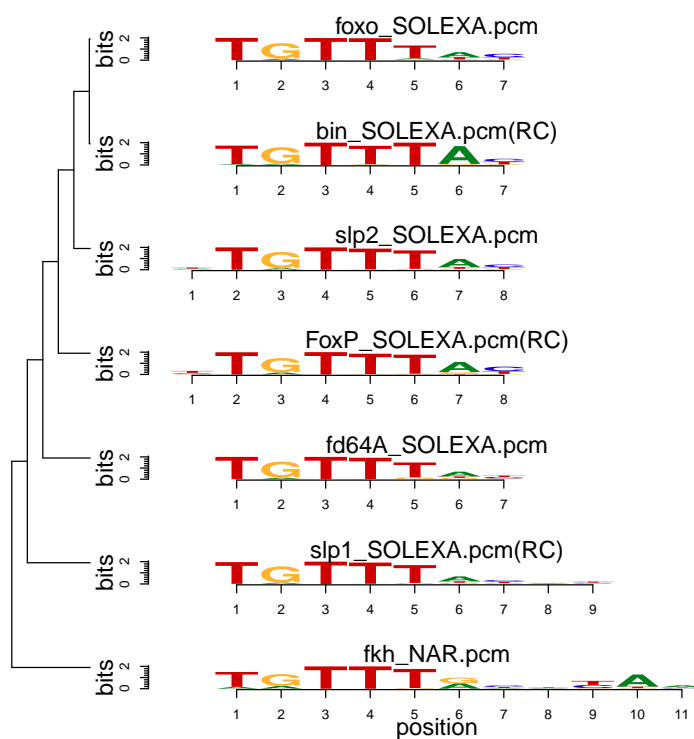


图 4: 带有关系树的图标堆

```

> pfms <- sample(motifs, 50)
> jasper <- MotIV::readPWMfile(file.path(find.package("MotIV"), "extdata", "jaspar2010.txt"))
> jasper.scores <- MotIV::readDBScores(file.path(find.package("MotIV"), "extdata", "jaspar2010_PCC_SWU.scores"))
> ffs.analysis.jasper <- MotIV::motifMatch(inputPWM=pfms, align="SWU",
+                                       cc="PCC", database=jaspar,
+                                       DBscores=jaspar.scores, top=5)

    Ungapped Alignment
    Scores read
    Database read
    Motif matches : 5

> d <- MotIV::motifDistances(getPWM(ffs.analysis.jasper))
> 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)
> pfms <- DNAmotifAlignment(pfms, minimalConsensus=3)
> library(RColorBrewer)
> color <- brewer.pal(12, "Set3")

```

```
> plotMotifStackWithRadialPhylog(phylog, pfms, circle=0.9, cleaves = 0.5, clabel.leaves = 0.7,
col.bg=rep(color, each=5), col.leaves=rep(color, each=5))
```

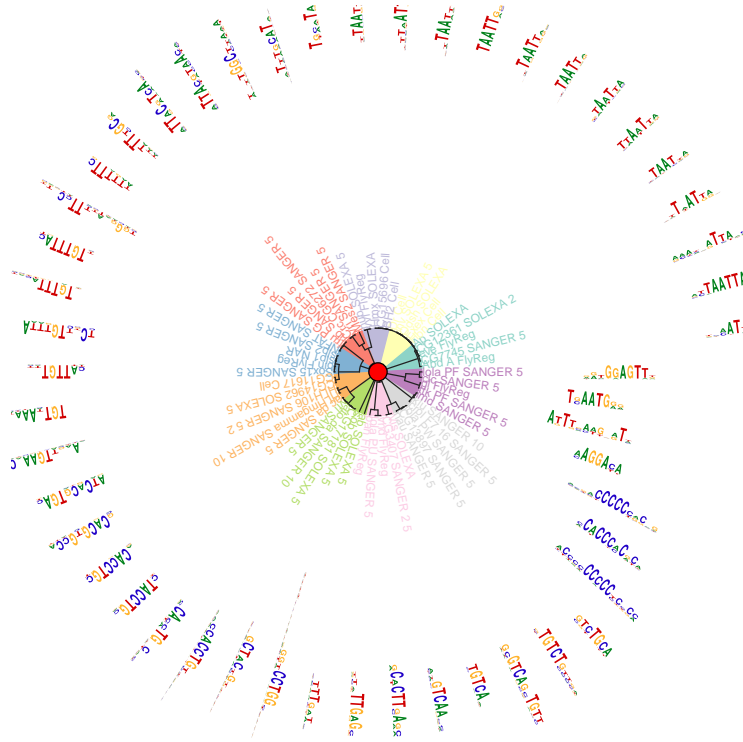


图 5: sequence logo stack in radial style

4 References

参考文献

- [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. Ma-

hony S, Benos PV, Nucleic Acids Res. 2007, 35(Web Server issue):
W253-W258.

5 Session Info

```
> sessionInfo()
```

```
R version 2.15.1 (2012-06-22)
```

```
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)
```

```
locale:
```

```
[1] en_US/en_US/en_US/C/en_US/en_US
```

```
attached base packages:
```

```
[1] grid      stats      graphics  grDevices  utils      datasets  methods
```

```
[8] base
```

```
other attached packages:
```

```
[1] RColorBrewer_1.0-5 MotifDb_1.0.0      Biostrings_2.26.2 IRanges_1.16.2
```

```
[5] MotIV_1.14.0      BiocGenerics_0.4.0 motifStack_1.2.3  ade4_1.4-17
```

```
[9] grImport_0.8-4    XML_3.95-0.1
```

```
loaded via a namespace (and not attached):
```

```
[1] bitops_1.0-4.1      BSgenome_1.26.1      GenomicRanges_1.10.2
```

```
[4] lattice_0.20-10    parallel_2.15.1      RCurl_1.95-1.1
```

```
[7] rGADEM_2.6.0       Rsamtools_1.10.1     rtracklayer_1.18.0
```

```
[10] seqLogo_1.24.0     stats4_2.15.1        tools_2.15.1
```

```
[13] zlibbioc_1.4.0
```