Feb 06, 2021

R editing of HMMSEARCH output: heatmap with associated with gene tree and protein illustration

Kanae Nishii[1,2]

[1]Royal Botanic Garden Edinburgh; [2]Kanagawa University

1 Works for me dx.doi.org/10.17504/protocols.io.bkqwkvxe

Kanae Nishii

ABSTRACT

R codes used for phyml tree and profile HMM heatmap visualization, protein domain illustration

DOI

dx.doi.org/10.17504/protocols.io.bkqwkvxe

PROTOCOL CITATION

Kanae Nishii 2021. R editing of HMMSEARCH output: heatmap with associated with gene tree and protein illustration. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bkqwkvxe

CREATED

Sep 03, 2020

LAST MODIFIED

Feb 06, 2021

PROTOCOL INTEGER ID

41462

R libraries

```
#libraries for tree drawing
library(ggtree)
library(ape)
library(phytools)

#libraries for heatmap
library(reshape2)
library(dplyr)
library(ggplot2)
library(aplot)
```

tree drawing

```
#read Newick format tree
tree1 <- read.tree("tree.nhx")
#check tree1
ggtree(tree1)
```

```
#rooting tree

#midpoint rooted tree
```

```
tree2 <- midpoint.root(tree1)

#rooted by outgroup samples
tree2 <- root(Mtree.outgroup=c("outgroup1", "outgroup2"), resolve.root=T)

#check tree2
ggtree(tree2)
```

4
```
#read data for grouping
dat <- read.csv(file="data.csv", header=T, na.strings=c("","-"))
dat2  <- data.frame(dat$group,dat$protein_ID)
dat2 <- dat2[!duplicated(dat2$dat.protein_ID),]
colnames(dat2) <- c("group","protein_ID")

#import groups
chlorophytes <- subset(dat2,dat2$group=="chlorophytes")
charophytes <- subset(dat2,dat2$group=="charophytes")
liverworts<- subset(dat2,dat2$group=="liverworts")
mosses<- subset(dat2,dat2$group=="mosses")
lycophytes<- subset(dat2,dat2$group=="lycophytes")
gymnosperms<- subset(dat2,dat2$group=="gymnosperms")
angiosperms<- subset(dat2,dat2$group=="angiosperms")
monocots<- subset(dat2,dat2$group=="monocots")
eudicots<- subset(dat2,dat2$group=="eudicots")

cls <- list(a.chrolophytes=chlorophytes$protein_ID,
b.charophytes=charophytes$protein_ID,
c.liverworts=liverworts$protein_ID,
d.mosses=mosses$protein_ID,
e.lycophytes=lycophytes$protein_ID,
f.gymnosperms=gymnosperms$protein_ID,
g.angiosperms=angiosperms$protein_ID,
h.monocots=monocots$protein_ID,
i.eudicots=eudicots$protein_ID)
```

5
```
#add group to tree
tree3 <- groupOTU(tree2,cls)
```

6
```
#tree drawing
obj <- ggtree(tree3) + aes(color=group) +
scale_color_manual(values=c("deepskyblue","green","seagreen",
"seagreen4","orange2","chocolate4","red","blue","gray40")) +
geom_tiplab() +
geom_text2(aes(subset=!isTip,label=label,hjust=1.5,vjust=-0.5))
ggsave("all_support_value.tree.pdf",obj, width=30, height=50, limitsize=FALSE)
```

7
```
#only includes support values > 0.7
q <- ggtree(tree3)
d <- q$data
d <- d[!d$isTip,]
d$label <- as.numeric(d$label)
d <- d[d$label > 0.7,]

#tree drawing
obj <- ggtree(tree3) + aes(color=group) +
scale_color_manual(values=c("deepskyblue","green","seagreen","seagreen4",
"orange2","chocolate4","red","blue","gray40")) +
geom_tiplab(size=6) +
geom_text(data=d,aes(label=label,hjust=1.5,vjust=-0.5,size=6))
```


Citation: Kanae Nishii (02/06/2021). R editing of HMMSEARCH output: heatmap with associated with gene tree and protein illustration. https://dx.doi.org/10.17504/protocols.io.bkqwkvxe

```
ggsave("strongsupport.tree.pdf",obj,width=60, height=90,limitsize=FALSE)

#done
```

Making heatmap from output of HMMER website search

```
#libraries
library(reshape2)
library(dplyr)
library(ggplot2)
library(aplot)
```

```
#generate data frame from "dat"
dat3 <- data.frame(dat$protein_ID,dat$hmm_name,dat$i.Evalue)
colnames(dat3) <- c("protein_ID","hmm_name","i.Evalue")

#remove duplicate, "i.Evalue" is ordered in the ascending manner
dat4 <- dat3[!duplicated(dat3[1:2]),]

#make table of i.Evalue, ordered by protein_ID and hmm_hame
dat5 <- dcast(dat4,dat4$protein_ID ~ dat4$hmm_name)

#format for ggtree
dat6 <- melt(dat5)
colnames(dat6) <- c("protein_ID","hmm_name","i.Evalue")
```

```
#order of domain, by preference
domorder <- c("PLAC8","MCAfunc","DUF2985","Pkinase_Tyr","UDPGT", "WI12","WRKY")
dat7 <- dat6 %>%
mutate(hmm_name=factor(hmm_name,levels=domorder),ordered=TRUE)
#dat7 is the data for heatmap
```

Combining tree and heatmap

```
tree2 #tree data
dat7 #heatmap data
```

```
#drawing
obj2 <- ggtree(tree2) + geom_tiplab(align = TRUE,size=0)
hm <- ggplot(dat7, aes(x=hmm_name,y=protein_ID)) +
geom_tile(aes(fill=i.Evalue),color="gray50") +
scale_fill_gradient(low="blue",high="yellow",na.value="gray80",limits=c(0,0.001)) +
scale_x_discrete(expand=c(1,0)) +
theme_tree2(axis.text.x=element_text(angle=70,vjust=0.5,hjust=0.5,size=8),
axis.text.y=element_text(size=8))
obj3 <- hm %>% insert_left(obj2)
ggsave("heatmap.pdf",obj3,width=10,height=50,limitsize=FALSE)
#done
```

Making protein domain illustration

```
df <- read.csv(file="HMMSEARCH_output.csv",header = T)
df$hmm_name <- as.character(df$hmm_name)
df$color <- as.character(df$color)
```

```
id <- df$ID
id <- unique(id)
```

```
for (i in 1:8){
df1 <- df[which(df$ID==id[i]),]
```


Citation: Kanae Nishii (02/06/2021). R editing of HMMSEARCH output: heatmap with associated with gene tree and protein illustration.
https://dx.doi.org/10.17504/protocols.io.bkqwkvxe

```
df1 <- df1[which(df1$bestdom==1),]
```

16
```
#set plot area
screen.width <- 1500
screen.height <- 25
protlength <- df1$length[1]

file_a <- paste(df1$protein_ID[1],"_dom.png", sep = "")
png(filea, width = 1500, height = 500 )
plot(c(-10, screen.width),
c(0,screen.height),
type = "n",
xlab = "Number of amino acids",
ylab = "", yaxt='n')
```

17
```
#make the protein frame
rect(xleft = 1,
ytop = screen.height/2+1.5,
ybottom = screen.height/2-1.5,
xright=protlength,
col="gray")
a <- length(df1$hmm_name)
for (i in 1:a){
rect(xleft=df1$ali_from[i],
ytop=screen.height/2+2.5,
ybottom=screen.height/2-2.5,
xright = df1$ali_to[i],
col= df1$color[i])
}
text(max(df1$length[1])/2, screen.height-2.5, df1$protein_ID, cex=1.5)
pos.text.x <- df1$ali_from[1:a] + (df1$ali_to[1:a] - df1$ali_from[1:a])/2
pos.text.y <- c(screen.height/2+3.5)
text (pos.text.x,pos.text.y,df1$hmm_name[1:a],cex = 1.5)
text(df1$length[1], screen.height/2-3,df1$length[1],cex=1.5)
text(1,screen.height/2-3,1,cex=1.5)
dev.off()
}

#done
```

18 Table 1. Example data format

group	protein_ID	ID	length	hmm_name	i.Evalue	ali_from	ali_to	num_doms	overlap	bestdom	color
charophytes	A0A1Y1HMH1_KLENI	A0A1Y1HMH1	986	MCAfunc	3.80E-17	13	150	3	FALSE	1	mistyrose
charophytes	A0A1Y1HMH1_KLENI	A0A1Y1HMH1	986	U-box	1.10E-20	254	325	3	FALSE	1	honeydew2
charophytes	A0A1Y1HMH1_KLENI	A0A1Y1HMH1	986	Arm	1.90E-07	735	772	3	FALSE	1	royalblue
charophytes	A0A1Y1HNZ6_KLENI	A0A1Y1HNZ6	1311	MCAfunc	4.10E-15	7	139	5	TRUE	1	mistyrose
charophytes	A0A1Y1HNZ6_KLENI	A0A1Y1HNZ6	1311	U-box	1.10E-20	246	315	5	TRUE	1	honeydew2
charophytes	A0A1Y1HNZ6_KLENI	A0A1Y1HNZ6	1311	Arm_2	1.30E-05	1014	1206	5	TRUE	0	royalblue4
charophytes	A0A1Y1HNZ6_KLENI	A0A1Y1HNZ6	1311	Arm	2.70E-06	1046	1085	5	TRUE	1	royalblue
charophytes	A0A1Y1HNZ6_KLENI	A0A1Y1HNZ6	1311	Arm	2.00E-07	1131	1169	5	TRUE	1	royalblue

Citation: Kanae Nishii (02/06/2021). R editing of HMMSEARCH output: heatmap with associated with gene tree and protein illustration. https://dx.doi.org/10.17504/protocols.io.bkqwkvxe