---

title: "Untitled"

author: "Nathan Brouwer"

date: "9/21/2021"

output: html\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

# A complete bioinformatics workflow in R

\*\*By\*\*: Nathan L. Brouwer

# "Worked example: Building a phylogeny in R"

## Introduction

### Vocab

## Software Preliminaries

### Download necessary packages

### Load packages into memory

```{r, message= F, warning=F}

# github packages

library(compbio4all)

# CRAN packages

library(rentrez)

library(seqinr)

library(ape)

# Bioconductor packages

library(msa)

library(Biostrings)

```

## XXXXXXXXing macromolecular sequences

```{r}

# Human shroom 3 (H. sapiens)

hShroom3 <- entrez\_fetch(db = "protein",

id = "NP\_065910",

rettype = "fasta")

```

```{r}

cat(hShroom3)

```

```{r}

# Mouse shroom 3a (M. musculus)

mShroom3a <- entrez\_fetch(db = "protein",

id = "AAF13269",

rettype = "fasta")

# Human shroom 2 (H. sapiens)

hShroom2 <- entrez\_fetch(db = "protein",

id = "CAA58534",

rettype = "fasta")

# Sea-urchin shroom

sShroom <- entrez\_fetch(db = "protein",

id = "XP\_783573",

rettype = "fasta")

```

```{r}

nchar(hShroom3)

nchar(mShroom3a)

nchar(sShroom)

nchar(hShroom2)

```

## Prepping macromolecular sequences

```{r}

fasta\_cleaner

```

```{r}

fasta\_cleaner <- function(fasta\_object, parse = TRUE){

fasta\_object <- sub("^(>)(.\*?)(\\n)(.\*)(\\n\\n)","\\4",fasta\_object)

fasta\_object <- gsub("\n", "", fasta\_object)

if(parse == TRUE){

fasta\_object <- stringr::str\_split(fasta\_object,

pattern = "",

simplify = FALSE)

}

return(fasta\_object[[1]])

}

```

```{r}

hShroom3 <- fasta\_cleaner(hShroom3, parse = F)

mShroom3a <- fasta\_cleaner(mShroom3a, parse = F)

hShroom2 <- fasta\_cleaner(hShroom2, parse = F)

sShroom <- fasta\_cleaner(sShroom, parse = F)

```

```{r}

hShroom3

```

## XXXXXXXXing sequences

```{r}

# add necessary function

align.h3.vs.m3a <- Biostrings:: (

hShroom3,

mShroom3a)

```

```{r}

align.h3.vs.m3a

```

```{r}

# add necessary function

Biostrings:: (align.h3.vs.m3a)

```

```{r}

align.h3.vs.h2 <- Biostrings::pairwiseAlignment(

hShroom3,

hShroom2)

```

```{r}

score(align.h3.vs.h2)

```

```{r}

Biostrings::pid(align.h3.vs.h2)

```

## The shroom family of genes

```{r}

shroom\_table <- c("CAA78718" , "X. laevis Apx" , "xShroom1",

"NP\_597713" , "H. sapiens APXL2" , "hShroom1",

"CAA58534" , "H. sapiens APXL", "hShroom2",

"ABD19518" , "M. musculus Apxl" , "mShroom2",

"AAF13269" , "M. musculus ShroomL" , "mShroom3a",

"AAF13270" , "M. musculus ShroomS" , "mShroom3b",

"NP\_065910", "H. sapiens Shroom" , "hShroom3",

"ABD59319" , "X. laevis Shroom-like", "xShroom3",

"NP\_065768", "H. sapiens KIAA1202" , "hShroom4a",

"AAK95579" , "H. sapiens SHAP-A" , "hShroom4b",

#"DQ435686" , "M. musculus KIAA1202" , "mShroom4",

"ABA81834" , "D. melanogaster Shroom", "dmShroom",

"EAA12598" , "A. gambiae Shroom", "agShroom",

"XP\_392427" , "A. mellifera Shroom" , "amShroom",

"XP\_783573" , "S. purpuratus Shroom" , "spShroom") #sea urchin

```

```{r}

# convert to XXXXXXXXXC

shroom\_table\_matrix <- matrix(shroom\_table,

byrow = T,

nrow = 14)

# convert to XXXXXXXXXC

shroom\_table <- data.frame(shroom\_table\_matrix,

stringsAsFactors = F)

# XXXXXXXXXC columns

names(shroom\_table) <- c("accession", "name.orig","name.new")

# Create simplified species names

shroom\_table$spp <- "Homo"

shroom\_table$spp[grep("laevis",shroom\_table$name.orig)] <- "Xenopus"

shroom\_table$spp[grep("musculus",shroom\_table$name.orig)] <- "Mus"

shroom\_table$spp[grep("melanogaster",shroom\_table$name.orig)] <- "Drosophila"

shroom\_table$spp[grep("gambiae",shroom\_table$name.orig)] <- "mosquito"

shroom\_table$spp[grep("mellifera",shroom\_table$name.orig)] <- "bee"

shroom\_table$spp[grep("purpuratus",shroom\_table$name.orig)] <- "sea urchin"

```

```{r}

shroom\_table

```

## XXXXXing multiple sequences

```{r}

shroom\_table$accession

```

```{r}

# add necessary function

shrooms <- (db = "protein",

id = shroom\_table$accession,

rettype = "fasta")

```

```{r, eval = F}

cat(shrooms)

```

```{r}

shrooms\_list <- entrez\_fetch\_list(db = "protein",

id = shroom\_table$accession,

rettype = "fasta")

```

```{r}

length(shrooms\_list)

```

```{r}

for(i in 1:length(shrooms\_list)){

shrooms\_list[[i]] <- fasta\_cleaner(shrooms\_list[[i]], parse = F)

}

```

```{r}

# XXXXXXXXCX

shrooms\_vector <- rep(NA, length(shrooms\_list))

# XXXXXXXXCX

for(i in 1:length(shrooms\_vector)){

shrooms\_vector[i] <- shrooms\_list[[i]]

}

# XXXXXXXXCX

names(shrooms\_vector) <- names(shrooms\_list)

```

```{r}

# add necessary function

shrooms\_vector\_ss <- Biostrings:: (shrooms\_vector)

```

## MSA

### Building an XXXXXXXX (MSA)

```{r}

# add necessary function

shrooms\_align <- (shrooms\_vector\_ss,

method = "ClustalW")

```

### Viewing an MSA

#### Viewing an MSA in R

```{r}

shrooms\_align

```

```{r}

class(shrooms\_align) <- "AAMultipleAlignment"

shrooms\_align\_seqinr <- msaConvert(shrooms\_align, type = "seqinr::alignment")

```

```{r, eval = F}

print\_msa(alignment = shrooms\_align\_seqinr,

chunksize = 60)

```

#### Displaying an MSA XXXXXXXX

```{r}

## add necessary function

ggmsa:: (shrooms\_align, # shrooms\_align, NOT shrooms\_align\_seqinr

start = 2000,

end = 2100)

```

#### Saving an MSA as PDF

```{r, eval = F}

msaPrettyPrint(shrooms\_align, # alignment

file = "shroom\_msa.pdf", # file name

y=c(2000, 2100), # range

askForOverwrite=FALSE)

```

```{r}

getwd()

```

## Genetic XXXXXXXXX

```{r}

### add necessary function

shrooms\_dist <- seqinr:: (shrooms\_align\_seqinr,

matrix = "identity")

```

```{r}

### add necessary function

shrooms\_dist\_rounded <- (shrooms\_dist,

digits = 3)

```

```{r}

shrooms\_dist\_rounded

```

## Phylognetic XXXXXX (finally!)

```{r}

### add necessary function

tree <- (shrooms\_dist)

```

### Plotting XXXXXX

```{r}

# plot tree

plot.phylo(tree, main="Phylogenetic Tree",

type = "unrooted",

use.edge.length = F)

# add label

mtext(text = "Shroom family gene tree - unrooted, no branch lengths")

```

```{r}

# plot tree

plot.phylo(tree, main="Phylogenetic Tree",

use.edge.length = F)

# add label

mtext(text = "Shroom family gene tree - rooted, no branch lenths")

```

```{r}

# plot tree

plot.phylo(tree, main="Phylogenetic Tree",

use.edge.length = T)

# add label

mtext(text = "Shroom family gene tree - rooted, with branch lenths")

```

```{r fig.width=6}

plot(tree, main="Phylogenetic Tree")

mtext(text = "Shroom family gene tree")

x <- 0.551

x2 <- 0.6

# label Shrm 3

segments(x0 = x, y0 = 1,

x1 = x, y1 = 4,

lwd=2)

text(x = x\*1.01, y = 2.5, "Shrm 3",adj = 0)

segments(x0 = x, y0 = 5,

x1 = x, y1 = 6,

lwd=2)

text(x = x\*1.01, y = 5.5, "Shrm 2",adj = 0)

segments(x0 = x, y0 = 7,

x1 = x, y1 = 9,

lwd=2)

text(x = x\*1.01, y = 8, "Shrm 1",adj = 0)

segments(x0 = x, y0 = 10,

x1 = x, y1 = 13,

lwd=2)

text(x = x\*1.01, y = 12, "Shrm ?",adj = 0)

segments(x0 = x, y0 = 14,

x1 = x, y1 = 15,

lwd=2)

text(x = x\*1.01, y = 14.5, "Shrm 4",adj = 0)

segments(x0 = x2, y0 = 1,

x1 = x2, y1 = 6,

lwd=2)

segments(x0 = x2, y0 = 7,

x1 = x2, y1 = 9,

lwd=2)

segments(x0 = x2, y0 = 10,

x1 = x2, y1 = 15,

lwd=2)

```