````
---
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)

```