Analiza Poravnav Mafft in Muscle algoritmov

```r
library(ggplot2)
library(patchwork)
library(gridExtra)
library(dplyr)
```

Nalaganje podatkov

```r
files <- list.files(path = getwd(), pattern = "tsv", recursive = TRUE)
selected_files <- files[grep(x = files, pattern = "muscle|mafft")]
selected_files
```

```
## [1] "mafft/mafft_rc_200PAM_stats.tsv"
## [2] "mafft/mafft_rc_20PAM_full_trim_stats.tsv"
## [3] "mafft/mafft_rc_20PAM_stats.tsv"
## [4] "muscle/msl_rc_full_trim_stats.tsv"
## [5] "muscle/msl_rc_stats.tsv"
```

```r
load_tsv <- function(file_path) {
  read.delim(file_path,
             header = TRUE,
             sep = "\t",
             stringsAsFactors = FALSE,
             dec = ".")
}
```

mafft dataseti:

```r
mafft <- list()

grep(x = selected_files, pattern = "mafft") |>
  sapply(\(x) {
    name <- selected_files[x]
    mafft[[name]] <<- load_tsv(name)
  })
```

```
##            [,1]         [,2]         [,3]
## SequenceID character,42 character,42 character,42
## Identity   numeric,42   numeric,42   numeric,42
## Coverage   numeric,42   numeric,42   numeric,42
## Mismatches integer,42   integer,42   integer,42
```

muscle dataseti:

```
muscle  <- list()

grep(x = selected_files, pattern =  "muscle") |>
  sapply(\(x) {
    name <- selected_files[x]
    muscle[[name]] <<- load_tsv(name)
  })
```

```
##            [,1]         [,2]
## SequenceID character,42 character,42
## Identity   numeric,42   numeric,42
## Coverage   numeric,42   numeric,42
## Mismatches integer,42   integer,42
```

> Datasets imajo 42 zaporedij, ker je zraven še `consensus` zaporedje.

Povzetki

```
summ_col <- function(df, col) {
  cat(col, "\n")
  col <- df[, col]
  print(c(Min    = min(col),
          Q1     = quantile(col, 0.25),
          Median = median(col),
          Q1     = quantile(col, 0.75),
          Max    = max(col)))
  cat("\n")
}
```

```
summ_col2 <- function(df, col) {
  col <- df[, col]
  c(Min    = min(col),
    Q1     = quantile(col, 0.25),
    Median = median(col),
    Q1     = quantile(col, 0.75),
    Max    = max(col))
}
```

```
cols <- c("Identity", "Coverage", "Mismatches")
```

Mafft

```
purrr::walk(names(mafft), \(x) {
  cat("file: ", x, "\n\n")
  d <- mafft[[x]]
  purrr::walk(cols, summ_col, df = d)
  cat("------------------------------------------\n\n")
})
```

```
## file:  mafft/mafft_rc_200PAM_stats.tsv
##
## Identity
##     Min  Q1.25%  Median  Q1.75%     Max
##  75.990  86.135  96.510  97.020 100.000
##
## Coverage
##     Min  Q1.25%  Median  Q1.75%     Max
##  68.240  70.905  80.715  85.785 100.000
##
## Mismatches
##    Min Q1.25% Median Q1.75%    Max
##   0.00  12.50  19.50 128.75 306.00
##
## ------------------------------------------
##
## file:  mafft/mafft_rc_20PAM_full_trim_stats.tsv
##
## Identity
##      Min   Q1.25%   Median   Q1.75%      Max
##  86.4700  98.6675  99.2200  99.4300 100.0000
##
## Coverage
##      Min   Q1.25%   Median   Q1.75%      Max
##  99.4300  99.9300  99.9300  99.9825 100.0000
##
## Mismatches
##    Min Q1.25% Median Q1.75%    Max
##   0.00   7.00  10.00  17.75 182.00
##
## ------------------------------------------
##
## file:  mafft/mafft_rc_20PAM_stats.tsv
##
## Identity
##    Min Q1.25% Median Q1.75%    Max
##  76.27  86.89  96.51  97.02 100.00
##
## Coverage
##     Min  Q1.25%  Median  Q1.75%     Max
##  68.240  70.905  80.715  85.785 100.000
##
## Mismatches
##    Min Q1.25% Median Q1.75%    Max
##   0.00  12.50  19.50 128.75 302.00
##
## ------------------------------------------
```

Razlike med 200PAM in 20PAM poravnavo

```
name_pam200 <- names(mafft) |> grep(pattern = "200PAM_stats")
name_pam20  <- names(mafft) |> grep(pattern = "20PAM_stats")
cat(names(mafft)[name_pam200], "-", names(mafft)[name_pam20], "\n\n")

purrr::walk(cols, \(x) {
  mafft_pam200 <- mafft[[name_pam200]][, x]
  mafft_pam20  <- mafft[[name_pam20]][, x]

  diff <- mean(mafft_pam200) - mean(mafft_pam20)
  cat(x, ":", diff, "\n")
})
```

```
## mafft/mafft_rc_200PAM_stats.tsv - mafft/mafft_rc_20PAM_stats.tsv
##
## Identity : -0.785
## Coverage : 0
## Mismatches : 0.1190476
```

Razlike med neprečiščeno 20PAM in prečiščeno 20PAM

```
name_pam20_clean <- names(mafft) |> grep(pattern = "full_trim")
name_pam20_raw   <- names(mafft) |> grep(pattern = "20PAM_stats")
cat(names(mafft)[name_pam20_clean], "-", names(mafft)[name_pam20_raw], "\n\n")

purrr::walk(cols, \(x) {
  mafft_pam20_clean <- mafft[[name_pam20_clean]][, x]
  mafft_pam20_raw   <- mafft[[name_pam20_raw]][, x]

  diff <- mean(mafft_pam20_clean) - mean(mafft_pam20_raw)
  cat(x, ":", diff, "\n")
})
```

```
## mafft/mafft_rc_20PAM_full_trim_stats.tsv - mafft/mafft_rc_20PAM_stats.tsv
##
## Identity : 6.547143
## Coverage : 20.10262
## Mismatches : -54.45238
```

Muscle

Razlike med neprečiščeno in prečiščeno Muscle poravnavo

```
purrr::walk(names(muscle), \(x) {
  cat("file: ", x, "\n\n")
  d <- muscle[[x]]
  purrr::walk(cols, summ_col, df = d)
  cat("------------------------------------------\n\n")
})
```

```
## file:  muscle/msl_rc_full_trim_stats.tsv
##
## Identity
##      Min   Q1.25%   Median   Q1.75%      Max
##  86.5200  98.3175  98.9400  99.0800 100.0000
##
## Coverage
##    Min Q1.25% Median Q1.75%    Max
##  99.01  99.57  99.57  99.65 100.00
##
## Mismatches
##    Min Q1.25% Median Q1.75%    Max
##      0      7     10     18    176
##
## ------------------------------------------
##
## file:  muscle/msl_rc_stats.tsv
##
## Identity
##      Min   Q1.25%   Median   Q1.75%      Max
##  79.3000  86.5325  95.5700  96.5750 100.0000
##
## Coverage
##     Min  Q1.25%  Median  Q1.75%     Max
##  67.970  70.625  80.405  85.455 100.000
##
## Mismatches
##    Min Q1.25% Median Q1.75%    Max
##      0     12     19    128    301
##
## ------------------------------------------
```

```
name_clean <- names(muscle) |> grep(pattern = "full_trim")
name_raw   <- names(muscle) |> grep(pattern = "msl_rc_stats.tsv")
cat(names(muscle)[name_clean], "-", names(muscle)[name_raw], "\n\n")

purrr::walk(cols, \(x) {
  msl_clean <- muscle[[name_clean]][, x]
  msl_raw   <- muscle[[name_raw]][, x]

  diff <- mean(msl_clean) - mean(msl_raw)
  cat(x, ":", diff, "\n")
})
```

```
## muscle/msl_rc_full_trim_stats.tsv - muscle/msl_rc_stats.tsv
##
## Identity : 6.461905
## Coverage : 20.05738
## Mismatches : -54.33333
```

Primerjava prečiščenih poravnav Mafft in Muscle

```
name_mafft_clean <- names(mafft) |> grep(pattern = "full_trim")
name_msl_clean <- names(muscle) |> grep(pattern = "full_trim")
cat(names(mafft)[name_mafft_clean], "-", names(muscle)[name_msl_clean], "\n\n")

purrr::walk(cols, \(x) {
  mafft_clean <- mafft[[name_mafft_clean]][, x]
  msl_clean   <- muscle[[name_msl_clean]][, x]

  diff <- mean(mafft_clean) - mean(msl_clean)
  cat(x, ":", diff, "\n")
})
```

```
## mafft/mafft_rc_20PAM_full_trim_stats.tsv - muscle/msl_rc_full_trim_stats.tsv
##
## Identity : 0.3135714
## Coverage : 0.3469048
## Mismatches : 0.3571429
```

Ohranjena in variabilna mesta

```
filepath <- list.files(path = getwd(),
                       pattern = "sites_stats.tsv",
                       recursive = TRUE)
sites <- load_tsv(filepath)
sites
```

```
##      C    V  Pi   S Length                Alignment
## 1 1337  482 177 300   2064               msl_rc_fas
## 2 1326  495 178 309   2056        mafft_rc_20PAM.fas
## 3 1117  287 114 173   1409      msl_rc_full_trim.fas
## 4  282 1122 154 968   1422 mafft_rc_20PAM_full_trim.fas
```

```
calc_percentage <- function(df, col, len_col) {
  col_vals <- select(df, all_of(col)) |> unlist()
  len_vals <- select(df, all_of(len_col)) |> unlist()
  sapply(seq_len(nrow(df)), \(x) {
    round((col_vals[x] / len_vals[x]) * 100, digits = 2)
  })
}
```

```
lc <- "Length"

sites |>
  mutate(
    C_per  = calc_percentage(sites, "C", lc),
    V_per  = calc_percentage(sites, "V", lc),
    Pi_per = calc_percentage(sites, "Pi", lc),
    S_per  = calc_percentage(sites, "S", lc)
  ) |>
  select(C_per, V_per, Pi_per, S_per, Alignment)
```

```
##   C_per V_per Pi_per S_per                    Alignment
## 1 64.78 23.35   8.58 14.53                   msl_rc_fas
## 2 64.49 24.08   8.66 15.03           mafft_rc_20PAM.fas
## 3 79.28 20.37   8.09 12.28         msl_rc_full_trim.fas
## 4 19.83 78.90  10.83 68.07 mafft_rc_20PAM_full_trim.fas
```