Final Project HY390.51

July 01 2023

Student: Mourouzidou Eleni
MSc Bioinformatics

Description

In order to find the rate of the exponential growth of SARS-Cov-2, we have produced a set of 10000 simulated datasets and stored them in a file named “ms_sim_final.out”.
Our observed dataset is stored in a file called “ms_obs_final.out” and consists of 50 lines. Each line represents a genomic sequence of a Sars-Cov-2 strain derived from a European population. The length of every row (sequence) represents the number of polymorphic positions in the genome, compared to a reference Sars-Cov-2 genome. The file is in a binary format of 0s and 1s, where 0 stands for similarity and 1 stands for mutation in the corresponding position.

Every dataset of the simulated file has been generated using a growth parameter, stored in the file “pars_final.txt”. These datasets have the same number of sequences (50) equal to the number of sequences in the observed dataset. However, they have various sequence lengths (polymorphic positions) due to the nature of genetic variation in SARS-Cov-2. While the virus spreads and replicates, mutations can occur in its genome that lead to genetic diversity among different viral strains.
The varying sequence lengths (polymorphic positions) among the simulated datasets represent different levels of genetic variation observed in different strains of the virus.
Using three statistical metrics K, W, Tajima’s D we analyzed both the observed and simulated data. We aim to compare the observed dataset with the simulated

datasets and identify the growth parameter values that best describe the observed data.

Analysis

Having stored the observed dataset in a file, we converted it into a matrix of 1 column and 50 rows. The simulated datasets were converted into individual matrices per genome sequence and stored in a list called sim_matrices. We also built functions to calculate the statistics K, W, Tajima's D for each dataset.

K - statistic

```
calculate_K <- function(matrix_data) {
  d <- 0
  n <- n_calc(matrix_data)
  combs <- (n * (n - 1)) / 2

  for (i in 1:(n - 1)) {
    seq1 <- matrix_data[i, ]
    for (j in (i + 1):n) {
      seq2 <- matrix_data[j, ]
      diff_count <- sum(utf8ToInt(seq1) != utf8ToInt(seq2))
      d <- d + diff_count
    }
  }

  k <- d / combs
  return(k)
}
```

W - statistic

```
#       calculate a1 function
a1_calc <- function(matrix_data) {
  a1 <- 0
  n <- n_calc(matrix_data)
  for (s in 1:(n - 1)) {
    a <- 1/s
    a1 <- a1 + a
  }
  return(a1)
}
#      calculate W statistic using a1 and S
W_calc <- function(matrix_data) {
  return(S_calc(matrix_data)/a1_calc(matrix_data))
}
```

D - statistic

```
#       function for a2 calculation
a2_calc <- function(matrix_data) {
  a2 <- 0
  for (s in 1:(n_calc(matrix_data) - 1)) {
    a <- 1/s**2
    a2 <- a2 + a
  }
  return(a2)
}

#      function for b1 calculation
b1_calc <- function(matrix_data) {
  n = n_calc(matrix_data)
  return((n + 1)/(3*(n-1)))
}

#      function for b2 calculation
b2_calc <- function(matrix_data){
  n = n_calc(matrix_data)
  return((2*(n**2 + n + 3))/(9*n*(n-1)))
}
```

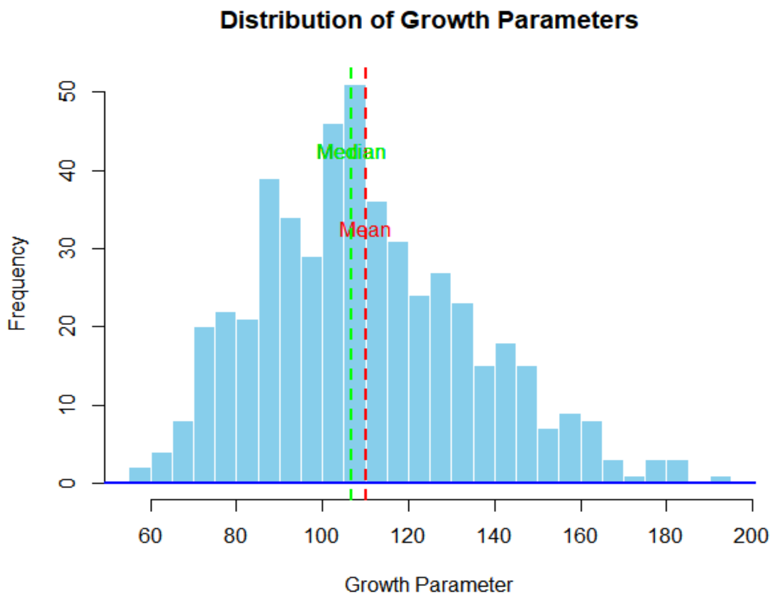
```r
#      function for c1 calculation
c1_calc <- function(matrix_data){
  b1 <- b1_calc(matrix_data)
  a1 <- a1_calc(matrix_data)
  return(b1 - 1/a1)
}

#       function for c2 calculation
c2_calc <- function(matrix_data){
  b2 <- b2_calc(matrix_data)
  a1 <- a1_calc(matrix_data)
  a2 <- a2_calc(matrix_data)
  n <- n_calc(matrix_data)
  return(b2 - ((n+2)/(a1*n)) + (a2/(a1^2)))
}

#       function for e1 calculation
e1_calc <- function(matrix_data){
  c1 <- c1_calc(matrix_data)
  a1 <- a1_calc(matrix_data)
  return(c1/a1)
}

#       function for e2 calculation
e2_calc <- function(matrix_data) {
  c2 <- c2_calc(matrix_data)
  a1 <- a1_calc(matrix_data)
  a2 <- a2_calc(matrix_data)
  return(c2/(a1^2 + a2))

}
#   Finally calculate Tajima's D
D_calc <- function(matrix_data){
  K <- calculate_K(matrix_data)
  W <- W_calc(matrix_data)
  e1 <- e1_calc(matrix_data)
  e2 <- e2_calc(matrix_data)
  S <- S_calc(matrix_data)
  return((K-W)/sqrt((e1*S)+ (e2*S)*(S-1)))

}
```

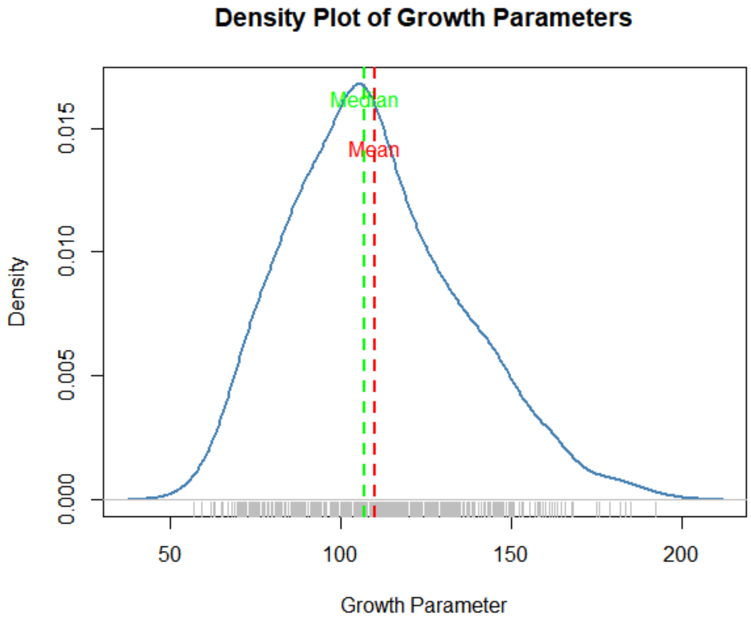
The observed dataset exhibited :
K = 14.28245, W = 29.46951, and Tajima's D = -1.845542.

We calculated the tree statistic both for the observed and simulated datasets and then normalized the statistic vectors by calculating the mean and standard deviation of the simulated datasets. We also normalized the observed statistic values using the same mean and standard deviation.

Euclidean distances were calculated between the normalized observed dataset and each of the normalized simulated datasets. We used a threshold of 500 to keep the smallest distances and their corresponding indexes were maintained.
The growth parameter values corresponding to the 500 smallest distances were retrieved from the “pars_final.txt” file. Finally we calculated the mean and median of these 500 growth parameters and created a histogram and density plot to visualize the distribution of the growth parameter values. These results are shown below:

```
> mean(best_growth)     #110.1053
[1] 110.1053
> median(best_growth)   #106.7937
[1] 106.7937
```





The mean growth parameter value of the 500 selected values was mean(best_growth) = 110.1053.
The median growth parameter value was median(best_growth) = 106.7937.

The distribution of the 500 growth parameters in the histogram plot, indicates that the majority of values are centered around 110 and the presence of more extreme values, indicates some level of variation in the growth rates of the virus population.

While exponential growth would typically provide a continuous increase in growth parameters, the observed distribution implies a different pattern.
The presence of values that deviate from the central tendency could indicate the influence of other factors in the growth rate of the virus. Factors that could influence the growth rate as shown in the results, could be genetic variations, or even healthcare, pharmaceutical, social or demographic factors.
It is also important to note that the distribution of the growth rate is apparently not exponential as we can also observe through the plots. The observed

distribution, suggests the mean value, 110.1053 represents the typical growth rate of the European Sars-Cov-2 population.

To conclude, observing the histogram, we could consider that the virus population is increasing by a factor of 110.1053 over a specific time period which indicates the virus spread trend. As we can observe from the density plot, the growth rate exhibits a single “sharp” peak which indicates a relatively consistent growth rate among the simulated datasets and as a result in the observed dataset. Considering that at least and the early stages of an epidemic an exponential growth is observed, we could conclude that this dataset is not derived from such a scenario. The growth distribution seems to follow a logistic distribution which is likely to happen when the growth rate of the virus may be influenced by factors such as the limited availability of individuals to get infected. Thus, the growth rate decays in this later stage, as more individuals become infected and the pool of susceptible individuals starts to diminish. This could be an explanation of the distribution that our observed dataset derives from. (Wu et al. ,2020)

References

Wu K, Darcet D, Wang Q, Sornette D. Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world. Nonlinear Dyn. 2020;101(3):1561-1581. doi: 10.1007/s11071-020-05862-6. Epub 2020 Aug 19. PMID: 32836822; PMCID: PMC7437112.