Stochastic Processes for Sequence Analysis

Assignment 2

Guillermo Carrillo Martín & Alberto Carrasco Parrón

2022-11-13

MSc in Bioinformatics Module 2

1. Introduction

The purpose of this exercise is to analyze the genome sequences of the **Zika virus** (**NC_012532.1**) and the **Dengue virus** (**NC_001477**).

First of all, the sequence data contained in the NCBI Nucleotide database is obtained by using the `rentrez` library.

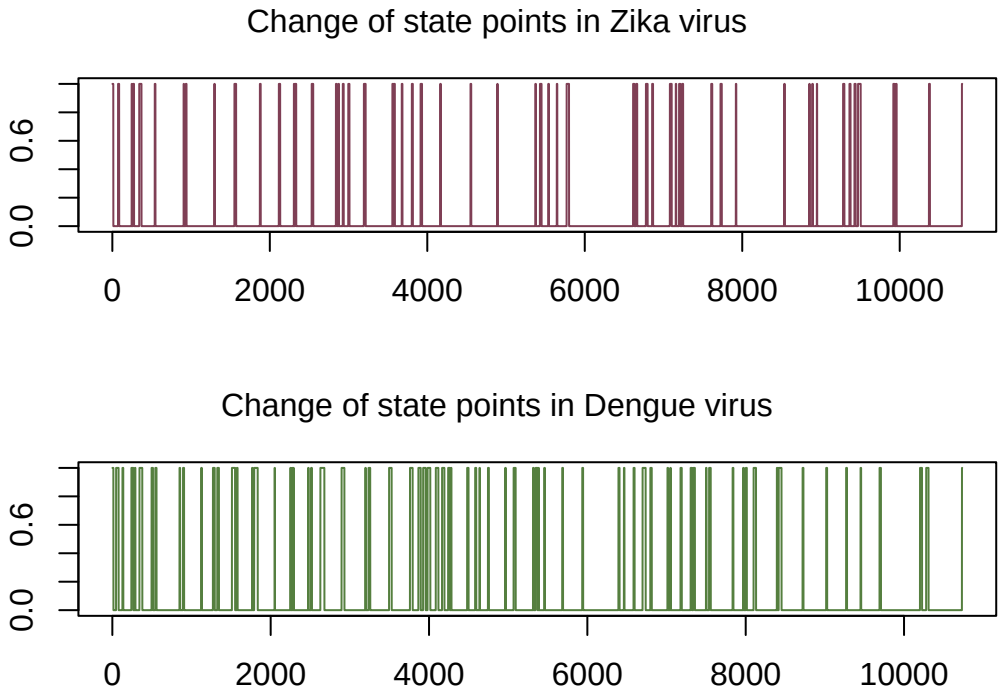
```
library(rentrez)
library(seqinr)
# Zika
zika_fasta <- rentrez::entrez_fetch(db = "nucleotide",
                                    id = "NC_012532.1",
                                    rettype = "fasta")
write(zika_fasta, file =  "input/zika.fasta")
zika <- read.fasta("input/zika.fasta")
zika <- zika[[1]]

# Dengue
dengue_fasta <- rentrez::entrez_fetch(db = "nucleotide",
                                      id = "NC_001477",
                                      rettype = "fasta")
write(dengue_fasta, file =  "input/dengue.fasta")
dengue <- read.fasta("input/dengue.fasta")
dengue <- dengue[[1]]
```

---

2. Hidden Markov chain Models (HMM)

2.1. Studying transition events

A HMM with two different states ("AT-rich" and "GC-rich") is built in order to infer which state is most likely to have generated each nucleotide in every sequence. Both the emission probabilities for every base under each state and the transition probabilities between states are known.

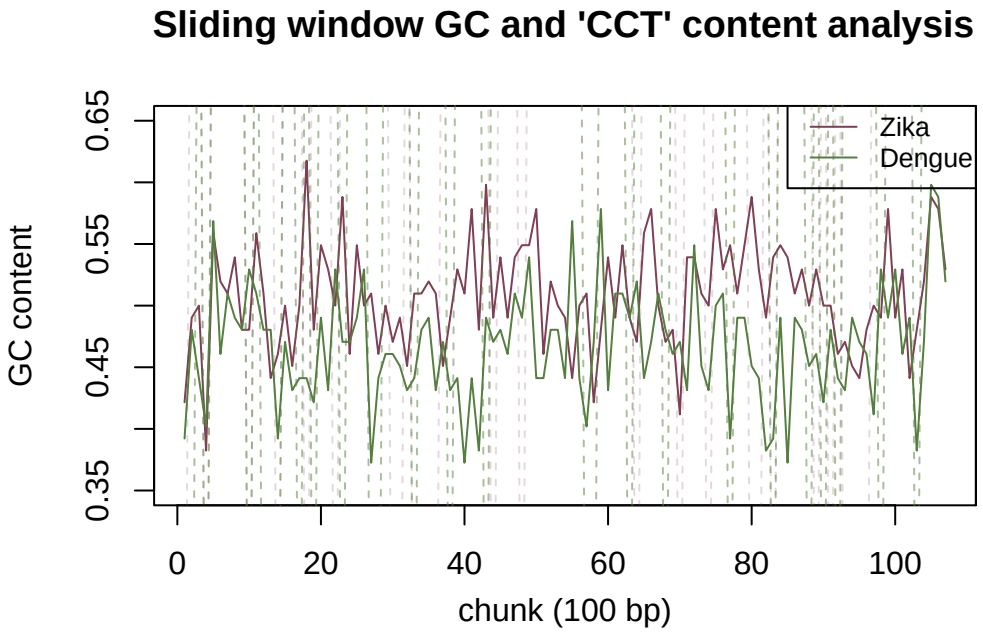


As it can be seen in the graphs, the amount of change points is **110** for Zika virus and **146** for Dengue. This means that the latter has more change of state points in its sequence.

2.2. Studying GC content and trinucleotid occurrence

The GC content of both sequences is calculated as well as the presence/absence of the trinucleotid CCT in chunks of length 100.

As seen in the graph, Zika virus shows a slight increase in the GC content when compared to Dengue. CCT occurrences are similarly distributed along both sequences. This trinucleotide appears across the whole length of the sequences, except for the mid part (chunks 50 to 60).



3. GLM

Let's build a simple logistic regression to discern whether there is any significant relationship between the presence of CCT and the GC content. This is the summary for the result of both linear models (one per virus):

```
## [1] "SUMMARY ZIKA"
```

```
##
## Call:
## glm(formula = cct_zika ~ gcc_zika, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2292   0.4746   0.6564   0.7756   1.0784
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.407      2.768  -1.592   0.1114
## gcc_zika      11.017      5.534   1.991   0.0465 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.66  on 106  degrees of freedom
## Residual deviance: 114.44  on 105  degrees of freedom
## AIC: 118.44
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "SUMMARY DENGUE"
```

```
##
## Call:
## glm(formula = cct_dengue ~ gcc_dengue, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2066   0.2522   0.5869   0.7359   1.3257
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.719      2.625  -2.560   0.0105 *
## gcc_dengue    17.116      5.787   2.958   0.0031 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118.66  on 106  degrees of freedom
## Residual deviance: 108.39  on 105  degrees of freedom
## AIC: 112.39
```

```
##
## Number of Fisher Scoring iterations: 4
```

The values for the independent variable in both GLM are significant - although more significant for Dengue virus. This means that there is indeed a significant relationship between GC content and CCT occurrences in these two sequences, and this association is stronger in Dengue (<0.001). One hypothesis behind this significant relationship is that it takes two Cs to form a CCT trinucleotide and therefore a higher GC content will provide them in a higher rate than AT-rich sequences.

Now, we can calculate the probability of the presence of CCT for a chunk with a GC content of 0.50 in each of the viruses. The estimation is possible using a logistic regression fit. The probability for Dengue is higher given the stronger association that exists between the two parameters (GC content and CCT presence).

```
# Zika
(exp(-4.407+11.017*0.5)) / (1+exp(-4.407+11.017*0.5))
```

```
## [1] 0.7505411
```

```
# Dengue
(exp(-6.719+17.116*0.5)) / (1+exp(-6.719+17.116*0.5))
```

```
## [1] 0.8628304
```

Complete code for this assignment is available at this repository.