Stochastic Processes for Sequence Analysis

Assignment 2

José María González Romero and Emiliano Navarro Garre

2022-11-10

1. Download Zika virus (NC_012532.1) and Dengue virus (NC_001477).

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

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

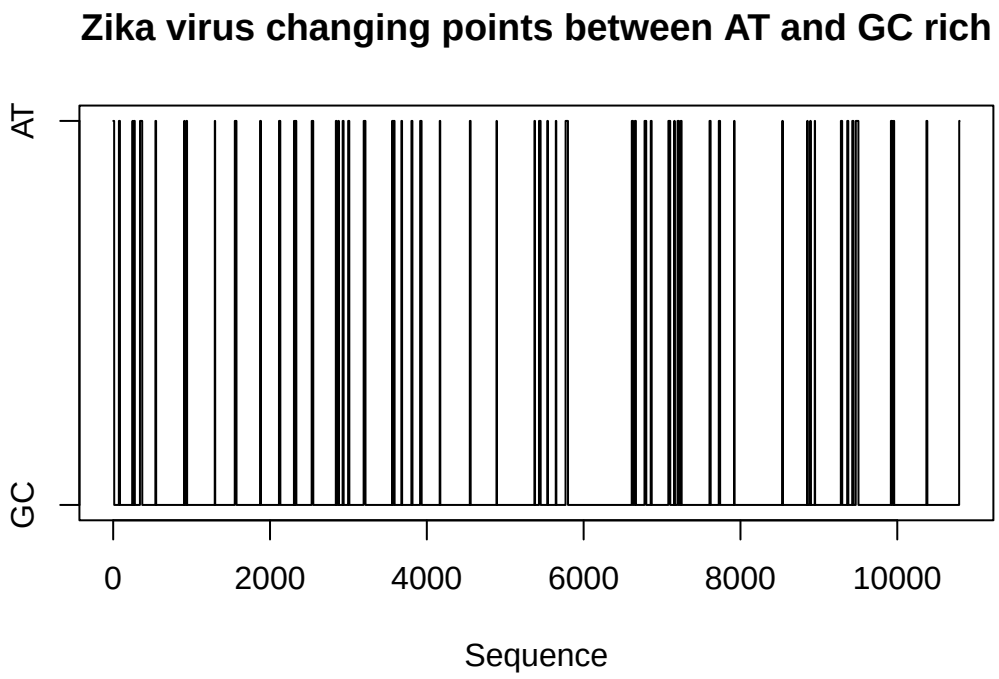
2. Some genomes have long stretches of either GC-rich or AT-rich sequence. Use a HMM with two different states ("AT-rich" and "GC-rich") to infer which state of the HMM is most likely to have generated each nucleotide position in Zika and Dengue sequences. In this case we exactly know the underlying HMM model, that is, for the AT-rich state, pA= 0.329, pC = 0.301, pG = 0.159, and pT = 0.211; for the GC-rich state, pA = 0.181, pC = 0.313, pG = 0.307, and pT = 0.199. Moreover, the probability of switching from the AT-rich state to the GC-rich state, or conversely, is 0.3. Make a plot for each virus in order to see the change points. Which of both viruses has more change points?

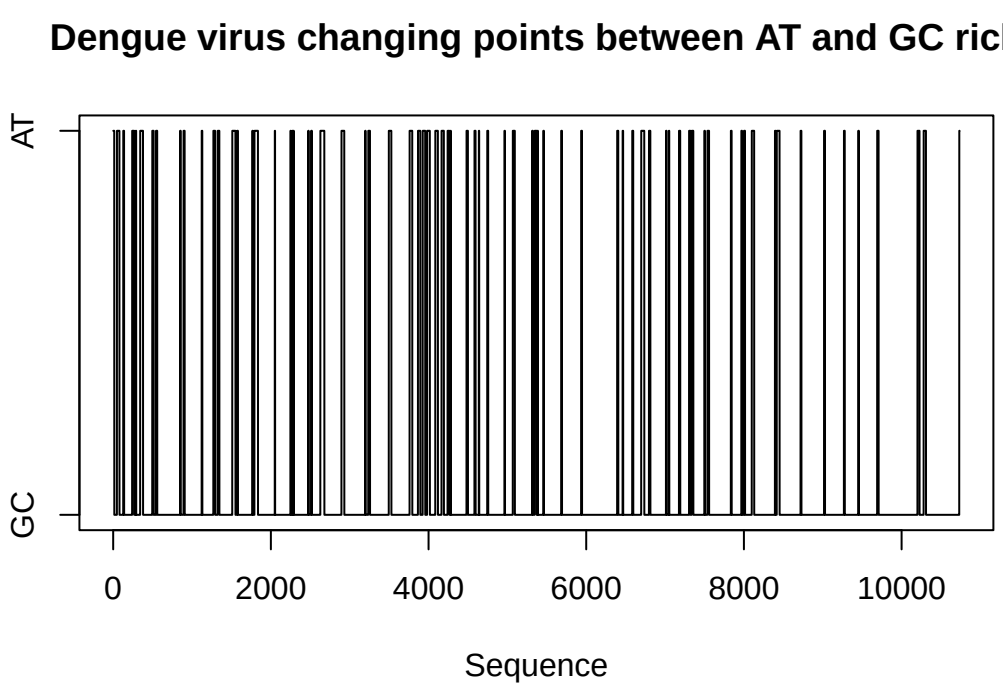
```
hmm=initHMM(c("AT","GC"), c("a","c","g","t"), c(0.5,0.5),
matrix(c(.7,.3,.3,.7),2), matrix(c(.329,.301,.159,.211,
.181,.313,.307,.199),2))

hmm
```

```
## $States
## [1] "AT" "GC"
##
```

```
## $Symbols
## [1] "a" "c" "g" "t"
##
## $startProbs
##  AT  GC
## 0.5 0.5
##
## $transProbs
##     to
## from  AT  GC
##   AT 0.7 0.3
##   GC 0.3 0.7
##
## $emissionProbs
##       symbols
## states     a     c     g     t
##     AT 0.329 0.159 0.181 0.307
##     GC 0.301 0.211 0.313 0.199
```

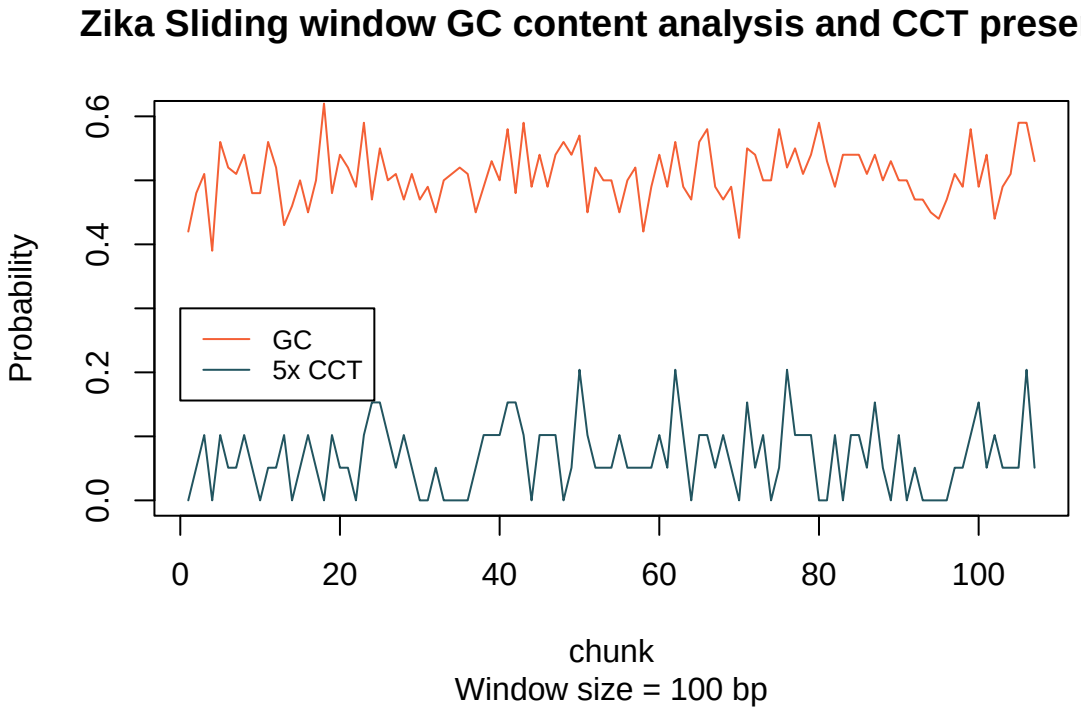
Plots for the changing points between AT and GC rich for both genomes are shown below.

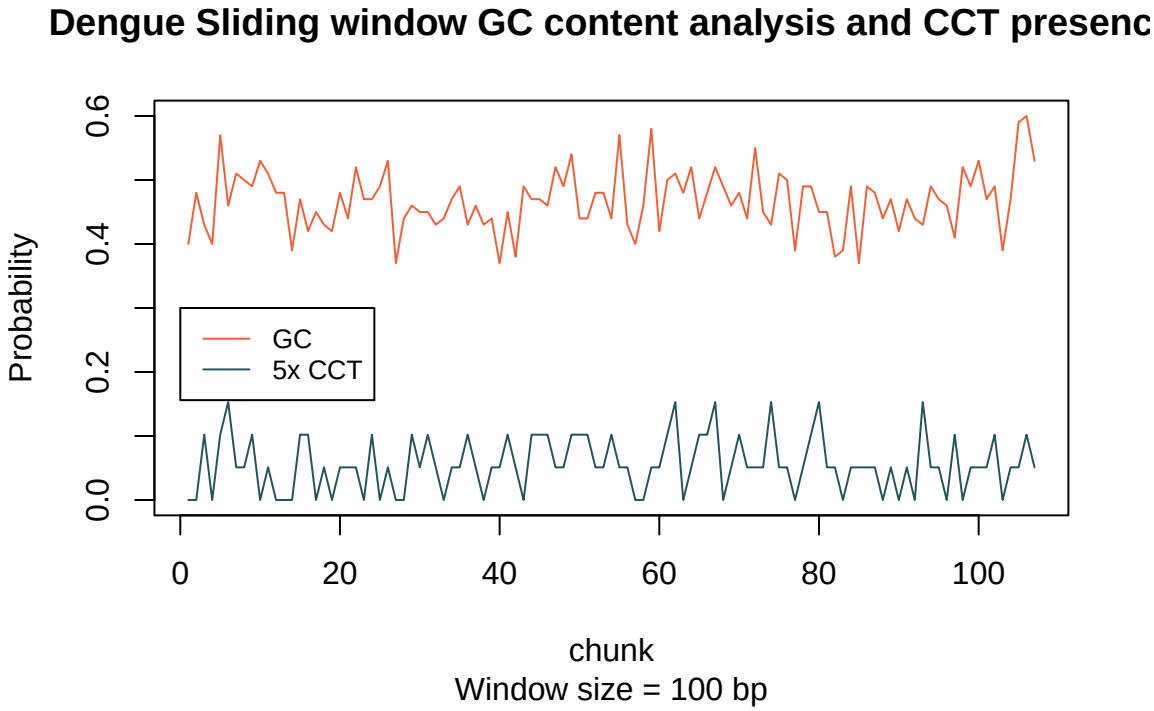




Between both viruses, Dengue virus has more changing points (146) from AT rich and GC rich, and conversely. Zika virus has 110 changes.

3. Calculate the GC content and the presence/absence of the trinucleotid "cct", of chunks with length 100 (for both viruses).





4. Is there any significant relationship between the presence of CCT and the GC content? Discuss and compare the results for both viruses.

```
pcctz =ifelse(cctz>0,1,0)
logitz =glm(pcctz~gcz,family=binomial)
summary(logitz)
```

```
##
## Call:
## glm(formula = pcctz ~ gcz, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2658   0.4716   0.6171   0.7776   1.1043
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.682      2.769  -1.691   0.0909 .
## gcz           11.563      5.539   2.088   0.0368 *
## ---
## 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: 113.99  on 105  degrees of freedom
## AIC: 117.99
##
## Number of Fisher Scoring iterations: 4
```

The summary shows a significant relationship (p-value<0.05) between the presence of CC and the GC content.

```
pcctd =ifelse(cctd>0,1,0)
logitd =glm(pcctd~gcd,family=binomial)
summary(logitd)
```

```
##
## Call:
## glm(formula = pcctd ~ gcd, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1187  -1.0985   0.6309   0.8017   1.3272
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.076      2.488  -2.442  0.01459 *
## gcd           15.487      5.471   2.831  0.00465 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.01  on 106  degrees of freedom
## Residual deviance: 113.71  on 105  degrees of freedom
```

```
## AIC: 117.71
##
## Number of Fisher Scoring iterations: 4
```

The summary shows a significant relationship (p-value<0.01) between the presence of CCT and the GC content. Both viruses show significant relationship, but Dengue virus shows a stronger one. We expected a significant results for both genomes, as increasing the number of G and C, trinucleotides containing these bases will increase. This relationship could be seen in the plots above too.

5. What is the probability of the presence of "cct" for a chunk with GC content of 0.50 in Zika virus? What is this probability for Dengue virus?

```
prob = 0.5
num = exp(coefficients(logitz)[1]+coefficients(logitz)[2]*prob)
probz = num/(1+num)
```

```
num = exp(coefficients(logitd)[1]+coefficients(logitd)[2]*prob)
probd = num/(1+num)
```

The probability of the presence of CCT for a chunk with GC content of 0.50 in Zika virus is 0.75 and the probability for Dengue virus is 0.841. The probability for Dengue virus is bigger, it could be related to the fact that the significance between GC content and presence of CCT is greater than in Zika.

6. Data availability

All the code used to carry out this analysis and to generate the present file is available in the following link:

- assignment2.Rmd