Compare Dengue (NC_001477) with Zika virus (NC_012532.1)

Edgar G. Nogales

11/02/2016

1 Introduction

They are both mosquito borne viruses spread especially by the Aedes Aegypti mosquito. Both have similar symptoms, including conjunctivitis, muscle and joint pain, rashes, headaches and fever.

http://www.toropest.com/which-is-the-difference-between-zika-and-dengue

Use all the techniques explained in the first lecture that you consider suitable. Please, comment the results and be creative! You can work in couples.

2 Exercise

For this evaluation we will use R as main software and the Seqinr library to perform the operations

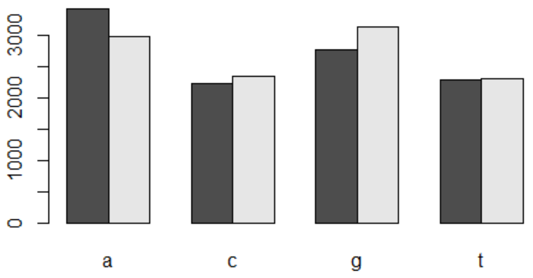
```
install.packages(”seqinr”)
library(”seqinr”)
```

We need to get the sequence in FASTA format and import it to work in the local environment.

```
zika <- read.fasta(file = ”zika.fasta”)
dengue <- read.fasta(file = ”dengue.fasta”)
```

First of all, we will calculate the frequency for each nucleotide, we will use barplot to compare the results.

```
dengu <- dengue[[1]]
freqDengue<-count(dengu,1)

zik <- zika[[1]]
freqZika<-count(zik,1)
```



Here we can compare the frequency, and we can see that Dengue has more "A" nucleotides and in the Zika virus the nucleotide with more repetition is the "G"
We need to check if both sequences have the same length (or at least similar).

```
length(zik)                             length(dengu)
[1] 10794                               [1] 10735
```

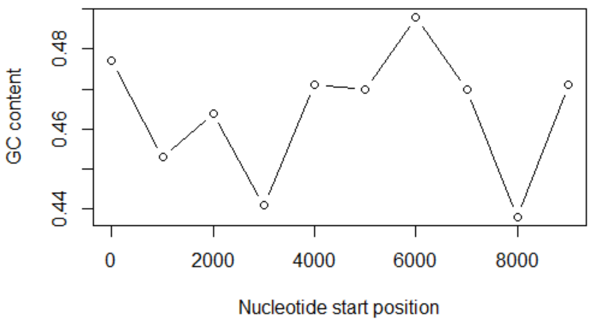
They are quite similar, this is a good point to start our study.
Next step will be the GC's content. This value is important because it has a relation with the coding regions, higher value will give us a more coding regions (genes).
We will use a SeqinR function to make our work easier and faster.

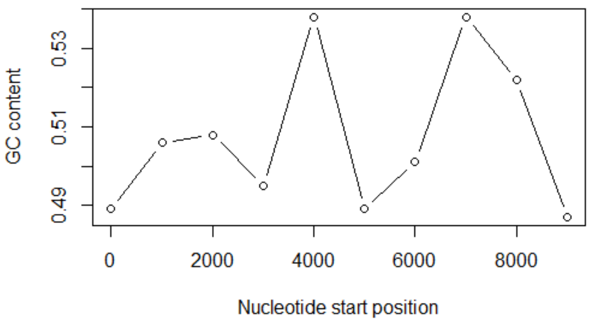
```
zikaGC<-GC(zik)

dengueGC<-GC(dengu)

> zikaGC                                          > dengueGC
[1] 0.5093571                                     [1] 0.4666977
```

With this values, we can assume that in the Zika virus we will find more coding regions.
It's possible to perform a study for the GC content using alternative windows slicing.
We have done a secondary function to make our job easy and more readable.

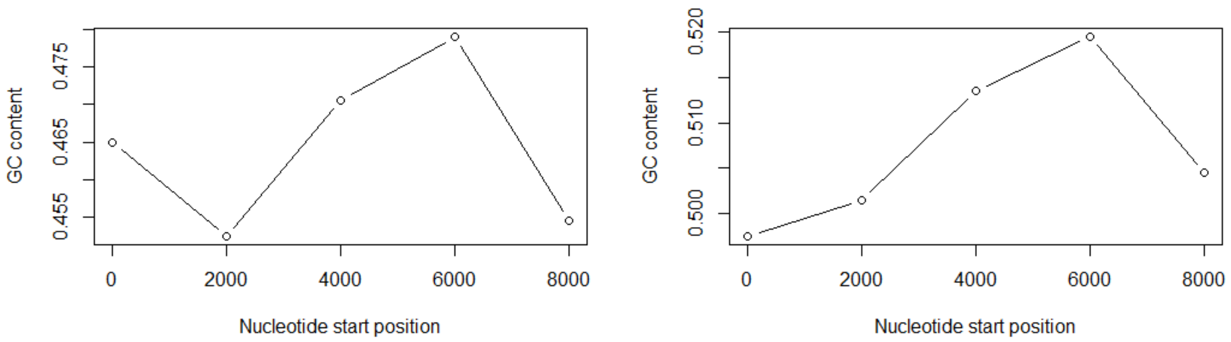
```
slidingwindowplot <- function(windowsize, inputseq)
{
  starts <- seq(1, length(inputseq)-windowsize, by = windowsize)
  n <- length(starts)
  chunkGCs <- numeric(n)
  for (i in 1:n) {
    chunk <- inputseq[starts[i]:(starts[i]+windowsize-1)]
    chunkGC <- GC(chunk)
    print(chunkGC)
    chunkGCs[i] <- chunkGC
  }
  plot(starts,chunkGCs,type="b",
  xlab="Nucleotide␣start␣position",ylab="GC␣content")
}
```

We have compared 3 sizes of windows slicing (1000,2000,3000)
Windows Slide = 1000 (dengue - zika)

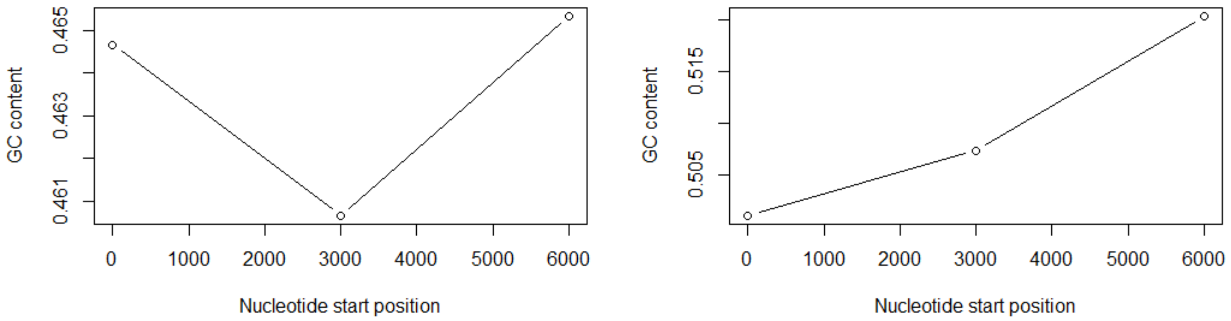




Windows Slide = 2000 (dengue - zika)



Windows Slide = 3000 (dengue - zika)



We can discover some differences. It's hard to predict how similar they are and find some similarities between both virus.
Next step will be to count the dinucleotides for each virus. We have 16 dinucleotides and we can estimate which are over or under-represented.

```
count(zik, 2)
aa   ac   ag   at   ca   cc   cg   ct   ga   gc   gg   gt   ta   tc   tg   tt
797  637  933  624  852  582  293  632 1003  626  951  559  338  514  962  490
```

```
count(dengu, 2)
aa   ac   ag   at   ca   cc   cg   ct   ga   gc   gg   gt   ta   tc   tg   tt
1108 720  890  708  901  523  261  555  976  500  787  507  440  497  832  529
```

We will try to find what kind of correlation exist between both virus. For that we will check first for 1 nucleotide, then for the dinucleotides and finally for groups of 3 bases.

```
zika1<-count(zik, 1)
dengue1<-count(dengu, 1)
cor(zika1,dengue1)
[1] 0.790663
```

```
zika2<-count(zik, 2)
dengue2<-count(dengu, 2)
cor(zika2,dengue2)
[1] 0.8688872
```

```
zika3<-count(zik, 3)
dengue3<-count(dengu, 3)
cor(zika3,dengue3)
[1] 0.8784204
```

We can see, that the correlation it's bigger for each step. It means that they are becoming more similar if we take bigger groups.

```
rho(zik)
aa        ac        ag        at
0.9617189 0.9745805 1.0727452 0.9770570
ca        cc        cg        ct
1.3035206 1.1289891 0.4271408 1.2547028
ga        gc        gg        gt
1.1532298 0.9125942 1.0418868 0.8340118
ta        tc        tg        tt
0.5292392 1.0204386 1.4352762 0.9955816
```

```
rho(dengu)
aa        ac        ag        at
1.0134622 1.0072555 1.0068514 0.9650492
ca        cc        cg        ct
1.2604683 1.1190466 0.4516013 1.1570403
ga        gc        gg        gt
1.1041427 0.8651367 1.1011784 0.8547355
ta        tc        tg        tt
0.5997481 1.0361244 1.4026428 1.0745342
```

We can use z-score to have a good point of view of under-over represented values.

```
zscore(zik,modele="base")
aa           ac           ag           at           ca           cc
-1.5245110   -0.8646824    2.9963725   -0.7689997   10.3247175    3.7478924
cg          ct           ga           gc           gg           gt
-20.1550379  7.2921150    6.3115335   -3.0752166    1.7844916   -5.7543807
ta           tc           tg           tt
-15.7788772   0.5851565   15.0899001   -0.1246432
```

```
zscore(dengu,modele="base")
aa           ac           ag           at           ca           cc
   0.6538017    0.2642849    0.2866112   -1.2942692    9.4877423    3.2523894
   cg           ct           ga           gc           gg           gt
   -17.2062694  4.3616997    4.3565375   -4.2314011    3.6457174   -4.6334962
   ta                 tc           tg           tt
     -14.8217801    1.0033335   12.8430797    2.1045484
```

To find if the values are significant or not, we need that their absolute values are greater than 2.
For the Zika virus we can find : AG, CA, CC, CG, CT, GA, GC, GT, TA and TG pairs as values significant.
In the other hand, we can find the results for the Zika: CA, CC, CG, CT, GA, GC, GG, GT, TA, TG and TT.
The correlation between both values is greater than 0.75. It's for that we can assume that they have a kind of correlation between each other.
Another point to compare is the frequency for the triplets. In this case we want to know the triplet with more repetitions.

```
max(count(zik, 3))
386              >>gga
```

```
max(count(dengu, 3))
388              >>aaa
```

We can see that the more frequent triplet it's not the same. That can have some relation with the codons of each virus.

3 Conclusions

I'm not an expert in biology. But after the test I can see some similarities between the viruses. With the basic operations it's hard to find the points in common. But After perform more advance techniques we can see a similarity between Dengue and Zika virus. It was hard to interpret the plots of the windows slicing, because they look completely different. But almost all the other test give us a positive result for the similarities.