Markov Models

Jan Izquierdo & David Marquez

2023-11-25

Setting libraries and importing data

```
library(seqinr)
```

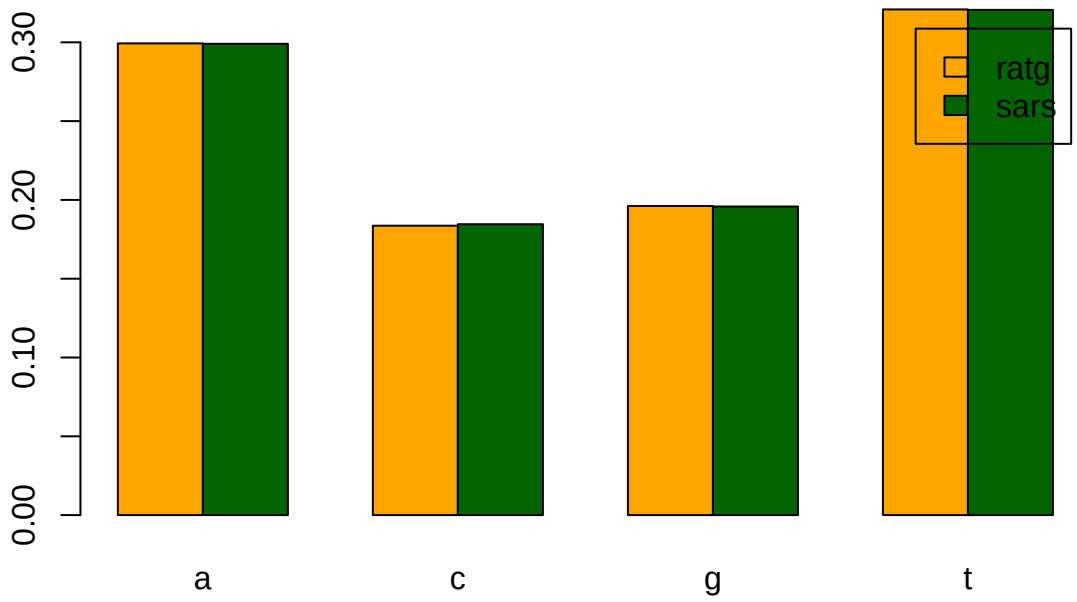
```
## Warning: package 'seqinr' was built under R version 4.2.3
```

1. **Download the genome of SARS-Cov-2 sequenced in Catalonia in 15/03/2020, accession number MT359865, and the genome of the virus RaTG13 of the bats Rhinolophus afinis, accession number MN996532.** .

```
s_data<-read.fasta("./assignment_data/SARSCov2.fasta")
r_data<-read.fasta("./assignment_data/RaTG13.fasta")
```

2. **Calculate the frequency of each nucleotide for each virus and draw Bar Plots. Which nucleotide is more frequent? Calculate the GC content and the AT content. Note that you have fitted the sequences to multinomial models: compute their BIC statistics.** .

```
sars<-s_data[[1]]
ratg<-r_data[[1]]
sars_freq<-table(sars)/length(sars)
ratg_freq<-table(ratg)/length(ratg)

GC_sars <- GC(sars)
GC_ratg <- GC(ratg)
AT_sars <- 1-GC_sars
AT_ratg <- 1-GC_ratg

combined_freq<-t(cbind(sars_freq, ratg_freq))

barplot(combined_freq, beside=T, col=c("orange","darkgreen"), legend.text = c("ratg", "sars"))
```



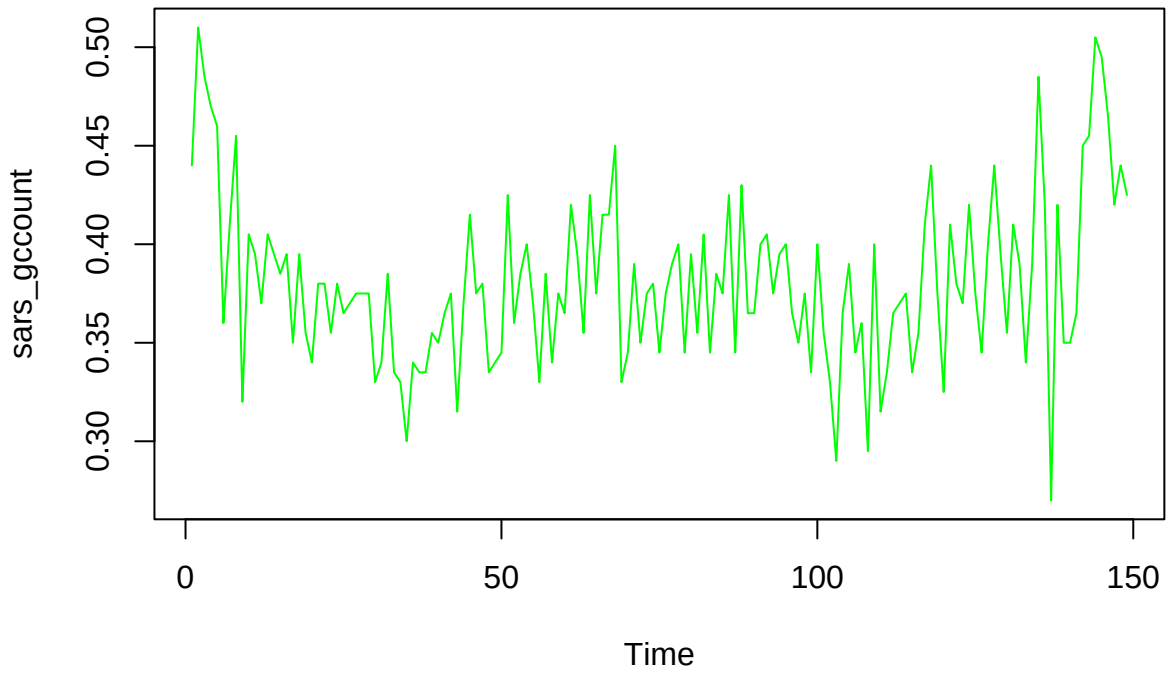

```
## The GC content of SARS-Cov-2 is 0.3797926

## The AT content of SARS-Cov-2 is 0.6202074

## The GC content if RaTG13 is 0.3803718

## The AT content if RaTG13 is 0.6196282
```

3. **Do a sliding window analysis of the GC content for each virus, that is, to study the variation in GC content within the genome sequence:** .

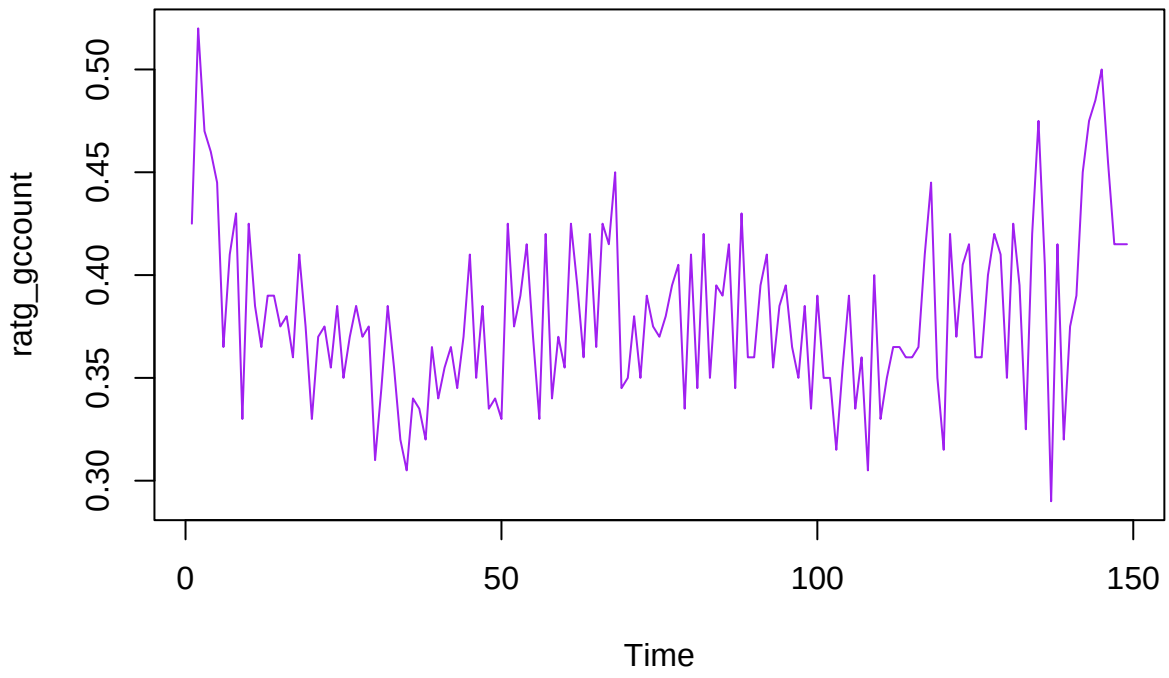
Calculate the GC content of chunks with length 200 (window size=200) and plot the resulting frequencies .

```
x=200
c=1
s=sars
n=length(sars)
y=200
sars_gccount=c()
i=1
while (x<=n){
    sars_gccount[i]<-GC(s[c:x]) #gc included function that calculates "g" "c" frequencies in chunk
```

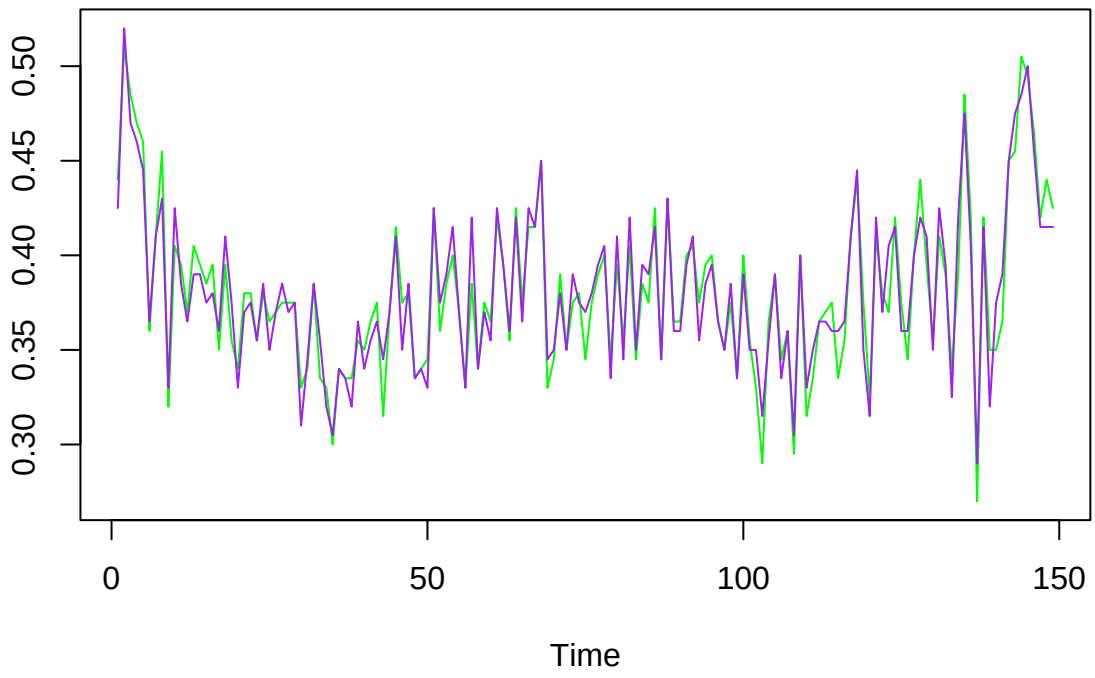
```
    c=c+y
    x=x+y
    i=i+1
}

x=200
c=1
s=ratg
n=length(ratg)
y=200
ratg_gccount=c()
i=1
while (x<=n){
    ratg_gccount[i]<-GC(s[c:x]) #gc included function that calculates "g" "c" frequencies in chunk
    c=c+y
    x=x+y
    i=i+1
}
ts.plot(sars_gccount, col="green")
```




```
ts.plot(ratg_gccount, col="purple")
```




```
combined_gcccount<-(cbind(sars_gccount, ratg_gccount))
ts.plot(combined_gcccount, col=c("green", "purple"))
```



***4.* Compute the number of occurrences of the dinucleotides for each virus. Which are over or under-represented?** .

```
s_c<-count(sars, 2)
r_c<-count(ratg, 2)
s_z<-zscore(sars, model="base")
r_z<-zscore(ratg, model="base")
```

```
## Dincucleotides for SARS-Cov-2:

##
##   aa   ac   ag   at   ca   cc   cg   ct   ga   gc   gg   gt   ta   tc   tg   tt
## 2878 2023 1740 2305 2084  887  438 2080 1611 1168 1094 1990 2373 1411 2591 3216

## Dinucleotides for RaTG13:

##
##   aa   ac   ag   at   ca   cc   cg   ct   ga   gc   gg   gt   ta   tc   tg   tt
## 2855 2037 1729 2306 2066  871  441 2132 1623 1161 1096 1966 2384 1440 2580 3167

##
##          aa          ac          ag          at          ca          cc
##   5.5152528  12.3968156  -0.4745183 -15.3083538  14.3866854  -4.6673136
```

```
##          cg          ct          ga          gc          gg          gt
## -24.0271924  10.2013541  -4.5777423   3.4369545  -2.0544844   3.3939744
##          ta          tc          tg          tt
## -13.4684933 -11.2083086  22.1481892   3.6783485
```

```
## AA, AC, CA, CT, GC, GT, TG and TT are over-represented, all others are under-represented
```

```
##
##          aa          ac          ag          at          ca          cc
##   5.1138578  12.6861592  -0.6103881 -15.0619656  13.6311485  -5.6103192
##          cg          ct          ga          gc          gg          gt
## -23.9823404  11.6887220  -3.9871389   3.0868662  -1.7893802   2.8731090
##          ta          tc          tg          tt
## -12.9492126 -10.4325505  22.0614148   2.6254669
```

```
## AA, AC, CA, CT, GC, GT, TG and TT are over-represented, all others are under-represented
```

5. **Fit both sequences to Markov chain models. Estimate their transition probability matrices. Compute the BIC of the models and find the steady-state (or limiting) distributions.** .

```
sars_m=matrix(s_c, 4, 4, byrow=T, dimnames=list(c("A", "C", "G", "T"), c("A", "C", "G", "T")))
sars_matrix<-sars_m[,]/(sars_m[,1]+sars_m[,2]+sars_m[,3]+sars_m[,4])
ratg_m=matrix(r_c, 4, 4, byrow=T, dimnames=list(c("A", "C", "G", "T"), c("A", "C", "G", "T")))
ratg_matrix<-ratg_m[,]/(ratg_m[,1]+ratg_m[,2]+ratg_m[,3]+ratg_m[,4])

sars_limiting_matrix <- sars_matrix%*%sars_matrix%*%sars_matrix%*%sars_matrix%*%sars_matrix%*%sars_matri

ratg_limiting_matrix <- ratg_matrix%*%ratg_matrix%*%ratg_matrix%*%ratg_matrix%*%ratg_matrix%*%ratg_matri
```

```
## The limiting distribution for SARS-Cov-2 is
```

```
##           A         C         G         T
## A 0.2993074 0.1836462 0.1961591 0.3208873
## C 0.2993074 0.1836462 0.1961591 0.3208873
## G 0.2993074 0.1836462 0.1961591 0.3208873
## T 0.2993074 0.1836462 0.1961591 0.3208873
```

```
## The limiting distribution for RaTG13 is
```

```
##           A         C         G         T
## A 0.2990538 0.1845338 0.1958229 0.3205895
## C 0.2990538 0.1845338 0.1958229 0.3205895
## G 0.2990538 0.1845338 0.1958229 0.3205895
## T 0.2990538 0.1845338 0.1958229 0.3205895
```

6. **Fit both sequences to second order Markov chain models, compute their BIC and compare the results with those obtained in 5. Which models are better?** .

```
sars_n=length(sars)-2; par=48
sars_c=sars_m
sars_p=sars_c[,]/(sars_c[,1]+sars_c[,2]+sars_c[,3]+sars_c[,4])
sars_BIC=-2*sum(sars_c*log(sars_p))+par*log(sars_n)

ratg_n=length(ratg)-2; par=48
ratg_c=ratg_m
ratg_p=ratg_c[,]/(ratg_c[,1]+ratg_c[,2]+ratg_c[,3]+ratg_c[,4])
ratg_BIC=-2*sum(ratg_c*log(ratg_p))+par*log(ratg_n)
```

```
## The BIC for SARS-Cov-2 is 80260.65
```

```
## The BIC for RaTG13 is 80205.16
```

7. Consider the following fragment of a DNA sequence: "g" "t" "g" "t" "g" "c" "t" "c" "a" "g" "t" "t" "g" "a" "a" "a" "a" "t" "c" "c" "c" "t" "t" "g" "t" "c" "a" "a" "c" "a" "t" "c" "t" "a" "g" "g" "t" "c" "t" "t" "a" "t" "c" "a" "c" "a" "t" "c" "a" "c" "a"

Decide using the log-likelihood method to which virus this sequence belongs, SARS-Cov_2 or RaTG13. .

```
u_seq <- c("g","t","g","t","g","c","t","c","a","g","t","t","g","a","a","a","a","t"
         ,"c","c","c","t","t","g","t","c","a","a","c","a","t","c","t","a","g",
         "g","t","c","t","t","a","t","c","a","c","a","t","c","a","c","a")
nucleotides<-c('a'=1, 'c'=2, 'g'=3, 't'=4)

res=c()
for (i in seq(length(u_seq)-1)){
  current_nuc<-nucleotides[u_seq[i]]
  f_nuc=nucleotides[u_seq[i+1]]
  res[i]=log(sars_matrix[current_nuc, f_nuc] / ratg_matrix[current_nuc, f_nuc])
}
res=(sum(res))
print(res)
```

```
## [1] -0.04506527
```

```
## The result of the log likelihood is -0.04506527
```

```
## Because in the formula we used the fraction in this format SARS-Cov-2/RaTG13 and -0.04506527 <0
## we can assume that the sequence does not belong to SARS-Cov-2, so it is from RaTG13
```