Names: Jakub Widawski & Xoel Mato Blanco

Stochastic processes for sequence analysis
Homework 1

Compare Dengue (NC_001477) with Zika virus (NC_012532.1). 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 lectures that you consider suitable.

a. Nucleotide frequencies

Genome composition analysis of nucleotide frequencies is known to be evolutionarily informative, and useful in metagenomic studies, where binning of raw sequence data is often an important first step. Patterns appearing in genome composition analysis may be due to evolutionary processes or purely mathematical relations.

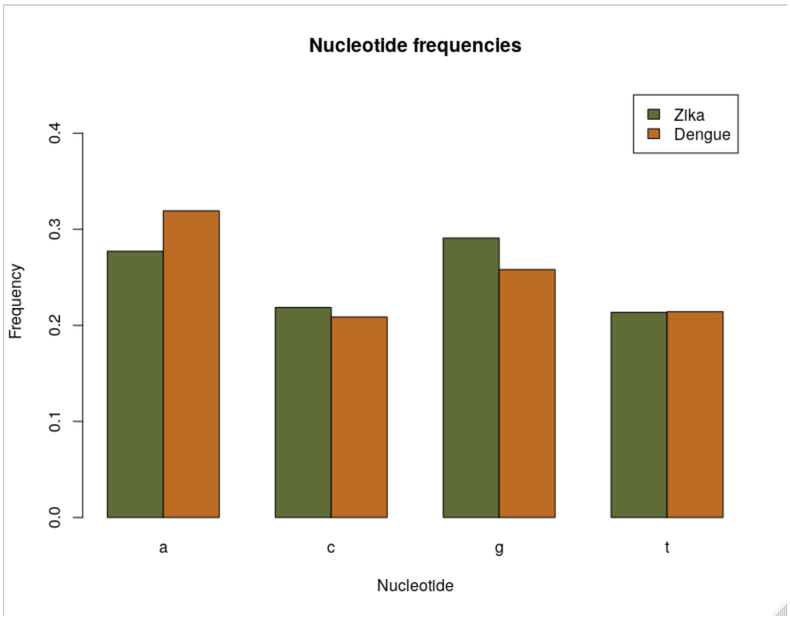


Figure 1. Nucleotide frequency comparison.

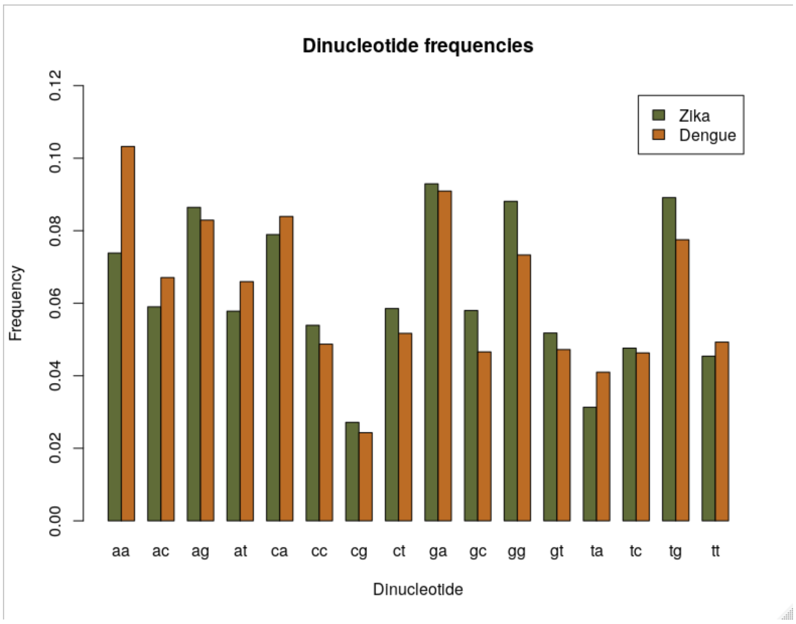


Figure 2. Dinucleotide frequency comparison.

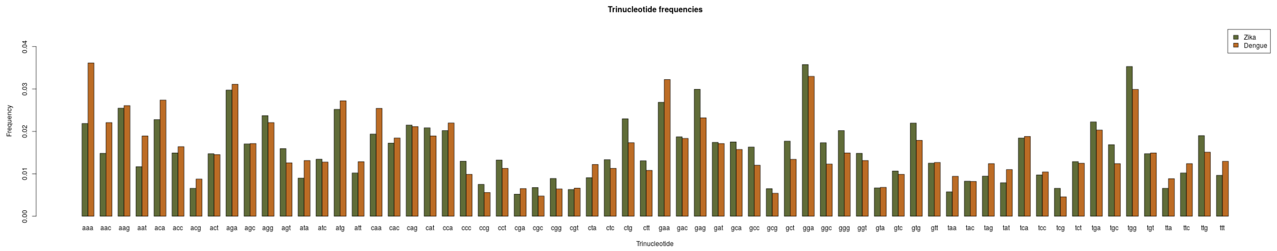
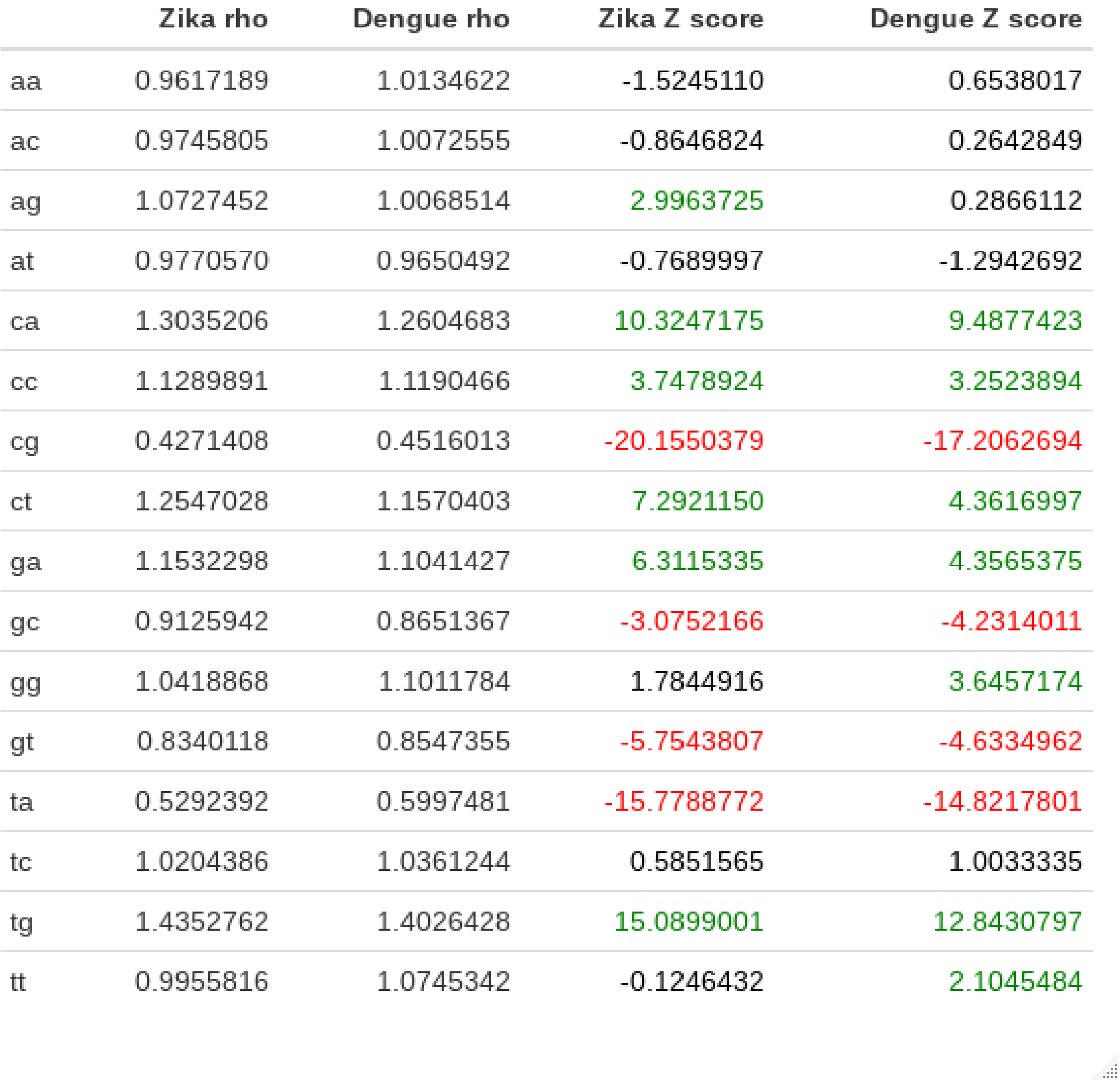


Figure 3. Trinucleotide frequency comparison. A bigger version of this figure is available as an attached file (Figure 3. Trinucleotide frequency comparison.png.)

b. Representation

The dinucleotide frequency is the incidence of a given neighbor dinucleotide in a sequence (e.g., a gene or a genome). When all nucleotides are used randomly (no selection), the frequencies of the sixteen dinucleotide pairs should be similar. However, studies have shown that several dinucleotide pairs can be over-presented or underrepresented in the genomes, suggesting the existence of selection pressure(s).



	Zika rho	Dengue rho	Zika Z score	Dengue Z score
aa	0.9617189	1.0134622	-1.5245110	0.6538017
ac	0.9745805	1.0072555	-0.8646824	0.2642849
ag	1.0727452	1.0068514	2.9963725	0.2866112
at	0.9770570	0.9650492	-0.7689997	-1.2942692
ca	1.3035206	1.2604683	10.3247175	9.4877423
cc	1.1289891	1.1190466	3.7478924	3.2523894
cg	0.4271408	0.4516013	-20.1550379	-17.2062694
ct	1.2547028	1.1570403	7.2921150	4.3616997
ga	1.1532298	1.1041427	6.3115335	4.3565375
gc	0.9125942	0.8651367	-3.0752166	-4.2314011
gg	1.0418868	1.1011784	1.7844916	3.6457174
gt	0.8340118	0.8547355	-5.7543807	-4.6334962
ta	0.5292392	0.5997481	-15.7788772	-14.8217801
tc	1.0204386	1.0361244	0.5851565	1.0033335
tg	1.4352762	1.4026428	15.0899001	12.8430797
tt	0.9955816	1.0745342	-0.1246432	2.1045484

Figure 4. Dinucleotide representation comparison.

We observed that the representation of dinucleotides is the same for both species except for two dinucleotides: AG is overrepresented only in Zika virus, and TT is overrepresented only in Dengue virus.

A comparison of trinucleotide representation can be found as an HTML file in the files attached.

c. GC content

i. Genome-wide

Genome-wide GC content is known to be very variable between species, and thus, it helps identifying different DNA samples. The following figures show that that is the case for these two viruses.

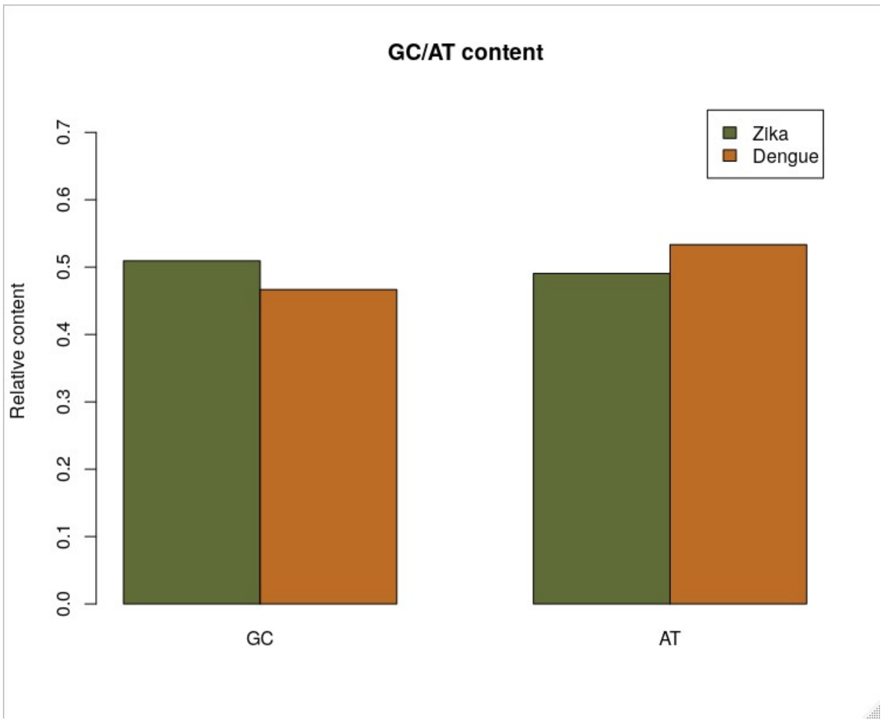


Figure 5. GC content comparison. GC content is higher in Zika virus compared to Dengue virus.

ii. Sliding windows

The GC content within a species' genome shows also great variability. Coding sequences usually have a greater GC content than non-coding sequences. Using sliding windows, we can infer which regions are more likely to be coding sequences.

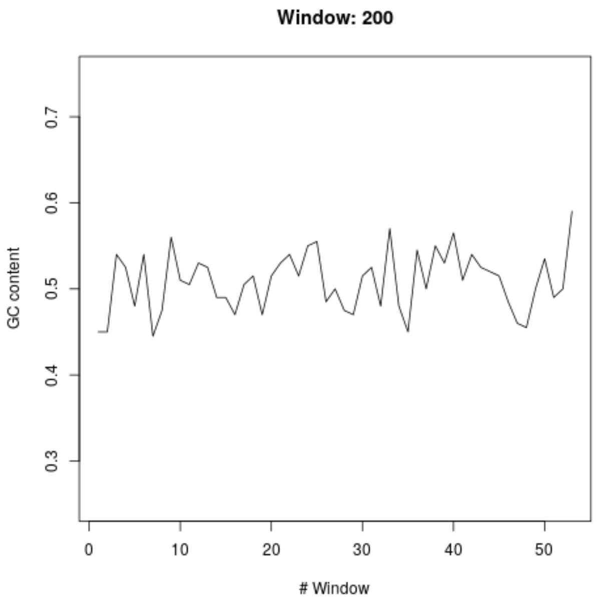


Figure 6A. Zika virus GC content.

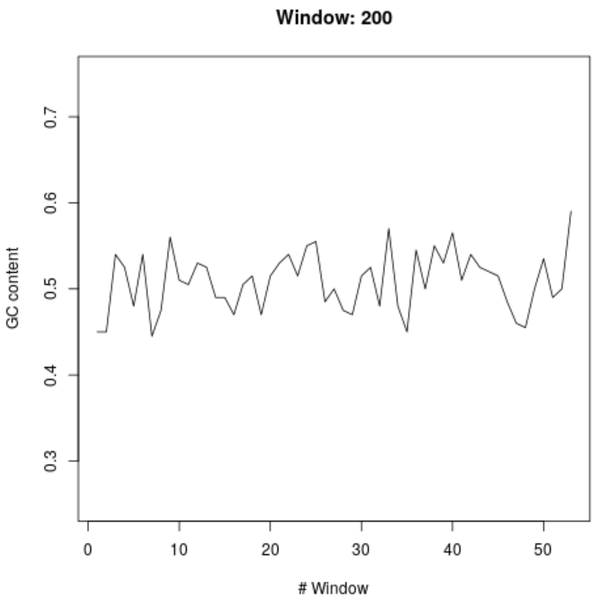


Figure 6B. Dengue virus GC content.

We did sliding windows of different lengths, but we chose to represent 200bp windows since they are the best visual representation for the GC content across the genome. The rest of the sliding window plots are available in the attached files.

d. Sliding windows for GpC representation

The dinculeotide CpG is partly responsible for the regulation of the gene expression in many organisms, and it also has a role in viral response. This dinucleotide is usually underrepresented across the genome, except for those regions that are regulated by cytosine methylation. Again, regions regulated by this mechanisms can be inferred studying CpG overrepresentation and sliding windows.

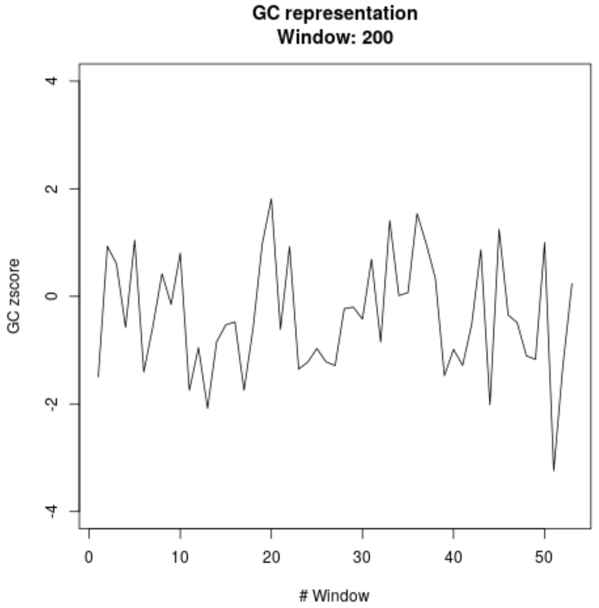


Figure 7A. Zika virus GC representation.

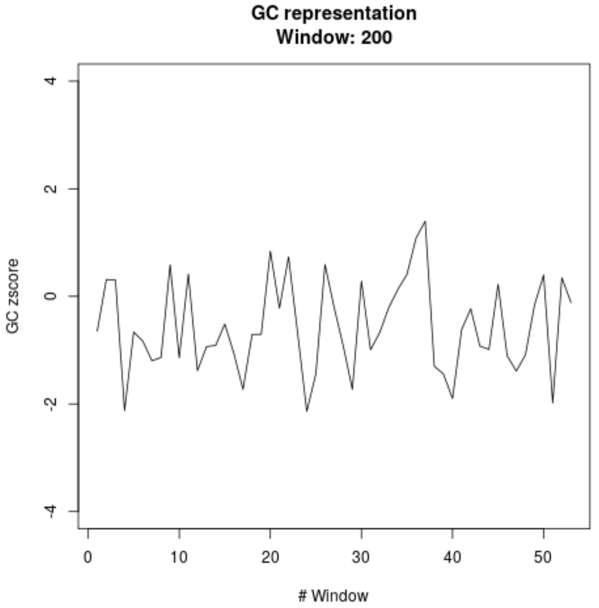


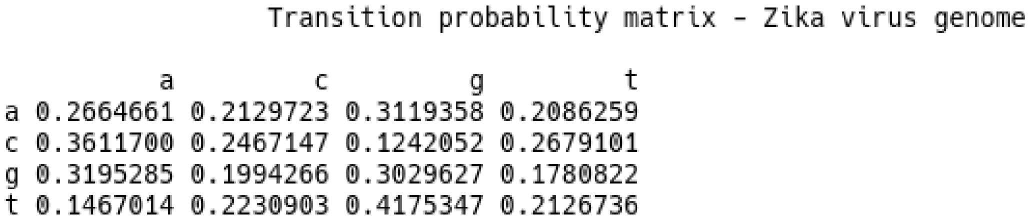
Figure 7B. Dengue virus GC representation.

Again, we chose a window length of 200bp since it is the best visual representation. The rest of the sliding window plots are available as attached files. A zscore greater/lower than +2/-1 is considered to be statistically significant for over/underrepresentation.

Question 2. Download Zika virus (NC_012532.1). Fit its genoma sequence to a Markov chain model, estimating its transition probability matrix.

a) Create a count matrix, transform it into a transition probability matrix

b) Transition probability matrix:

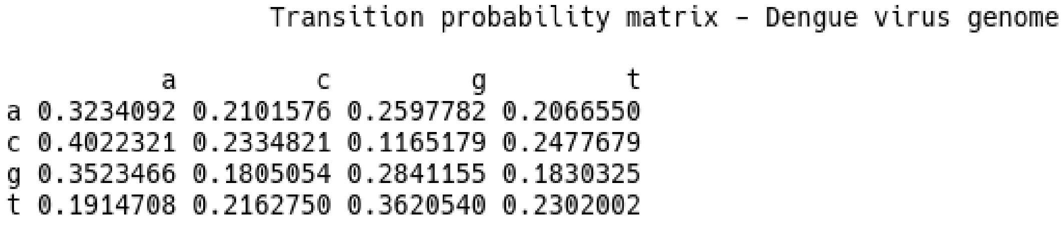


```
                  Transition probability matrix - Zika virus genome

          a         c         g         t
a 0.2664661 0.2129723 0.3119358 0.2086259
c 0.3611700 0.2467147 0.1242052 0.2679101
g 0.3195285 0.1994266 0.3029627 0.1780822
t 0.1467014 0.2230903 0.4175347 0.2126736
```

Question 3. Take the sequence of Dengue virus (NC_001477) from position 101 to 200 (this is a chunk of length 100). Suppose now that you don't know whether this sequence belongs to Zica or Dengue virus (of course, you know it!). Decide using the loglikelihood method to which virus this sequence belongs.

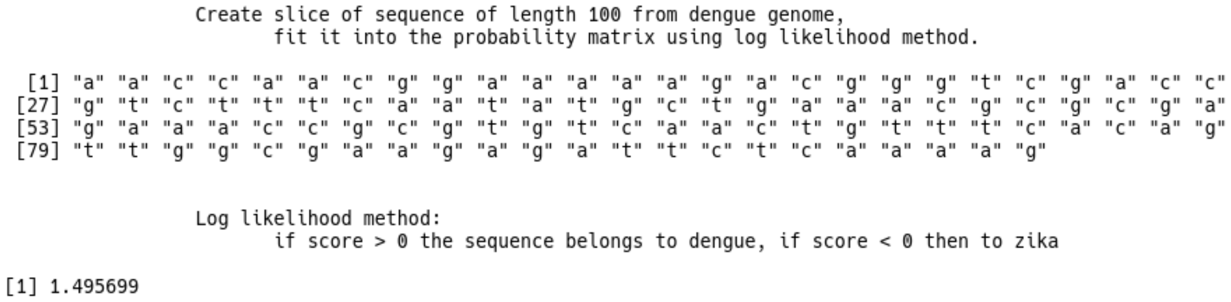
a) Create a count matrix, transform it into a transition probability matrix

b) Transition probability matrix:



```
                  Transition probability matrix - Dengue virus genome

          a         c         g         t
a 0.3234092 0.2101576 0.2597782 0.2066550
c 0.4022321 0.2334821 0.1165179 0.2477679
g 0.3523466 0.1805054 0.2841155 0.1830325
t 0.1914708 0.2162750 0.3620540 0.2302002
```

c) Create sample from positions 101-200, length = 100. Use log likelihood method to determine whether sequence belongs to dengue or zika.

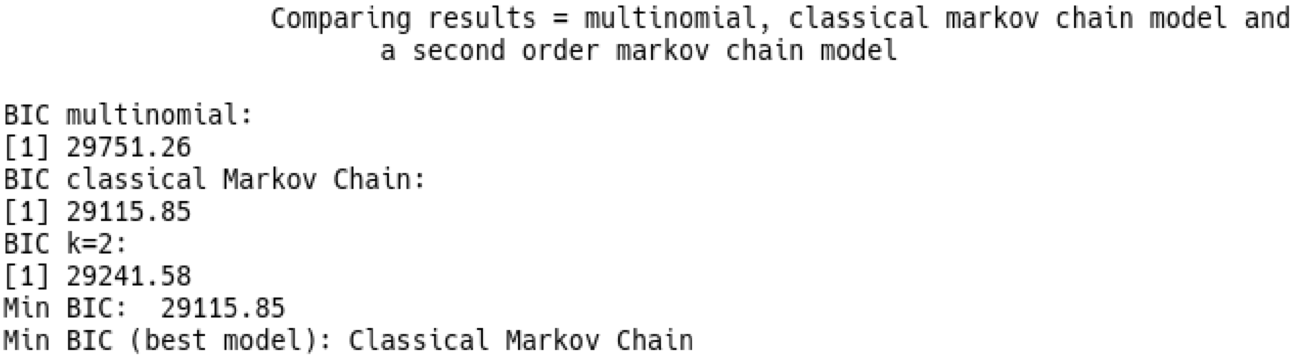


```
                Create slice of sequence of length 100 from dengue genome,
                      fit it into the probability matrix using log likelihood method.

  [1] "a" "a" "c" "c" "a" "a" "c" "g" "g" "a" "a" "a" "a" "a" "g" "a" "c" "g" "g" "g" "t" "c" "g" "a" "c" "c"
 [27] "g" "t" "c" "t" "t" "t" "c" "a" "a" "t" "a" "t" "g" "c" "t" "g" "a" "a" "a" "c" "g" "c" "g" "c" "g" "a"
 [53] "g" "a" "a" "a" "c" "c" "g" "c" "g" "t" "g" "t" "c" "a" "a" "c" "t" "g" "t" "t" "t" "c" "a" "c" "a" "g"
 [79] "t" "t" "g" "g" "c" "g" "a" "a" "g" "a" "g" "a" "t" "t" "c" "t" "c" "a" "a" "a" "a" "g"


                Log likelihood method:
                      if score > 0 the sequence belongs to dengue, if score < 0 then to zika

[1] 1.495699
```

Question 4. Fit the Zica virus sequence to a two second order Markov chain model. Compare the results with respect a simple Markov chain model.

Bayesian information criterion (BIC) is a criterion for model selection among a set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function.

a) Compare models using BIC



```
                Comparing results = multinomial, classical markov chain model and
                       a second order markov chain model

BIC multinomial:
[1] 29751.26
BIC classical Markov Chain:
[1] 29115.85
BIC k=2:
[1] 29241.58
Min BIC:  29115.85
Min BIC (best model): Classical Markov Chain
```