Hongzhe Li

SID 24474061

Topic: Identify the origin of replication in virus DNA

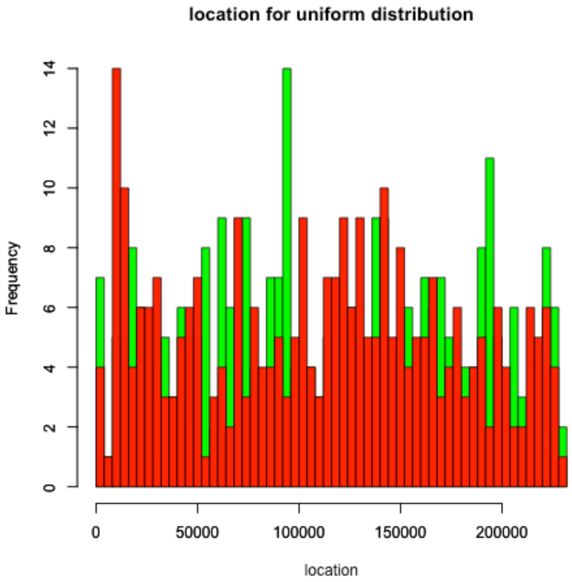
Introduction

Virus can infect all types of creatures, such as animals, plants, or even virus. As a result, to develop strategies for combating the virus, scientists study the way in which the virus replicates. In addition, virus replicates only inside of its cell; therefore, scientists are in search of a special place on the virus' DNA that contains instructions for its reproduction. Specifically in palindrome, which is a sequence of letters that reads in reverse as the complement of the forward sequence. In other words, the purpose of this lab is to find the origin of replication.

Since DNA is made of 2 long series of complicated strands, due to its numerous combinations and large length, there's a hardly possible way to find its pattern. Hence, I will cut DNA strands into many segments, and test each segments whether it can replicate or not, in order to find the origin of replication.

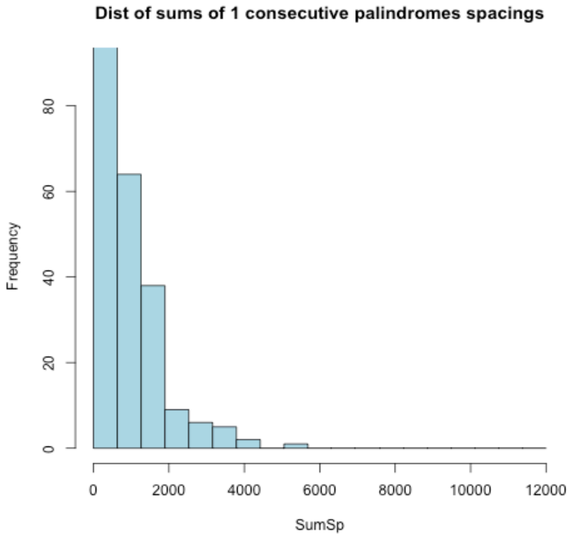
Methodology

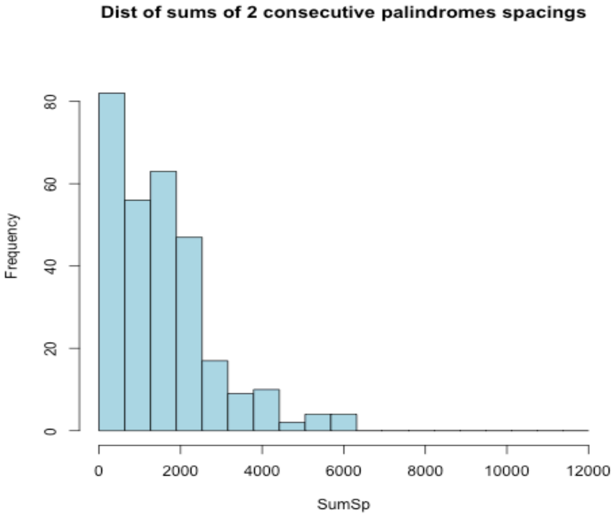
First of all, use the uniform generator function provided by R to simulate 1000 times of random uniform distribution with 229354 letters long, base (length of each interval) 4000, and calculate the mean. Thereafter, plot the histogram of both data given by the book, and the data we generated through R function in order to detect if the data follows the uniform distribution. The graph on the left is the histogram plotted according to the two stated data.



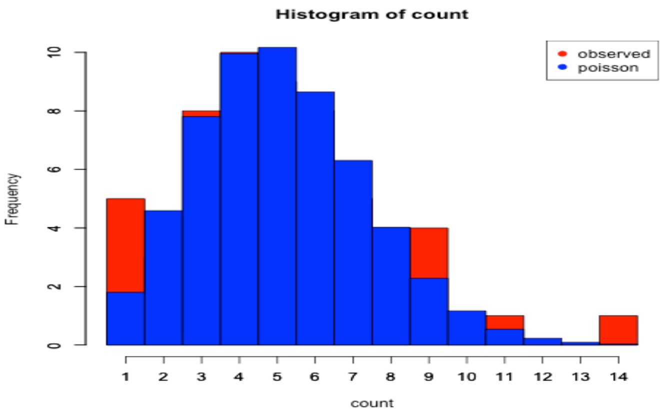
From the graph, it's obvious that these two graphs don't match each other. Therefore, the location of palindromes doesn't follow the uniform distribution.

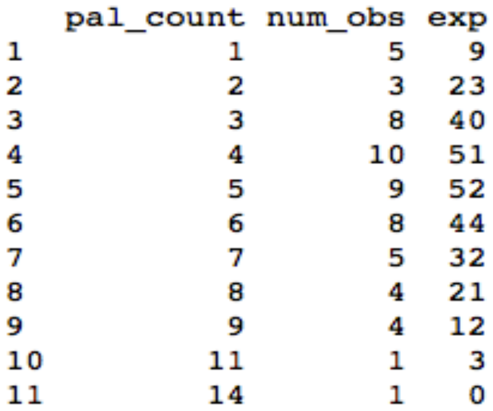
Second, we will discover whether the distances between successive hits follow the gamma distribution. By calculating the difference of each successive location, and plot the distance. The following is the graph, by analyzing the graph, and then calculate the chi-square to check if the space follows the gamma distribution. Via calculation, the p value I got is almost 0(actually, got the test of statistic of 276 which is not even on the table of chi-square). I will reject the null hypothesis, since the p-value is smaller than 0.05. Hence, it follows the gamma distribution.





Third, I will gather the counts of each interval by using the location of the real data, and examine whether it follows a Poisson distribution. I calculated the mean of counts for the real data as the rate (lambda) of the Poisson distribution. Formalize the table of the expected value of number of hits according to the data in each interval; then, graph the histogram and bar plot of the real data and Poisson distribution generated by the mean of the real data. Compare those 2 graphs




```
   pal_count num_obs exp
1          1       5   9
2          2       3  23
3          3       8  40
4          4      10  51
5          5       9  52
6          6       8  44
7          7       5  32
8          8       4  21
9          9       4  12
10        11       1   3
11        14       1   0
```

Last, in order to examine if the greatest number of hits in a collection of intervals behaves as the maximum of independent Poisson random variables by using the formula of maximum number of hits, I will calculate the p-value, from which I got the p-value(0.01610181), and it is less than 0.05. Therefore, I will reject the null hypothesis. In other words, the probability of the occurrence of the unusual count is very small.

Result

By analyzing the procedure stated above, I managed to find the origin of replicate is on the number 24th interval, and it is not a chance occurrence, but a potential replication site, since the location doesn't follow the uniform distribution, and the space and count follows the gamma and Poisson distribution.