DNA Pattern are potential replication sites

Lu Xu and Zehui(Barry) Zhang

Author Contributions

Zehui(Barry) Zhang and Lu Xu did the homework together and pretty equally. Both contributed to the doc writing, Lu did a little bit more, and complemented each other when one of us finished the first draft. We exchange ideas all the time. For the coding staff, mostly done by Barry, but Lu also came up with ideas all the time. We talk to each other all the time, and complement each other.

Index

1. Introduction…… 1
2. Basic Analysis…… 2
   2.1 Random scatter……..……. 3
   2.2 Locations and Spacings…….. 5
   2.3 Counts…… ……7
   2.4 The biggest cluster……..……..……8
3. Advanced Analysis ……9
4. Conclusion and Discussion ……...10

1.1 Introduction

The scientists want to build strong strategies to fight with the virus so they can study the replication of the virus. There are two parts of the virus' structures, DNA and RNA. They have a capsid which is their molecule coved the protein shell. DNA and RNA are a starting point of the viruses because it can be the genetic material to almost every virus. Both of them have their own replication and they also store most important information to contribute to our processes of life. DNA coding can create a complementary palindrome. This is like a pattern of the alphabet letter A, C, G, ant T with the complement reverse sequence. Also, DNA important patterns are called the origin of replication. Complementary meaning that one is complementary with another. For this study, it uses the human cytomegalovirus also known as CMV, or the herpes virus member. This virus will impact the geography a lot, to be more pacifice it is around 30 percent to 80 percent. One human being has more than 3 billions base pairs of DNA and CMV are including 229,354 of it. Herpes simplex and epstein-barr virus are also the members of herpes. 296 palindromes are included in CMV DNA and its base pair is around 10 to 18. In order to find out the origin of replication, it needs to be separated into segments. If those segments are replicated then the origins of replication are continued in it and opposite otherwise. Therefore, the statistical finds the better way to search those unusual clusters by narrowing down the testing amount of the complementary palindromes.

The goal of this analysis is to check whether the cluster is due by chance and compare the locations of those palindromes structured by chance to those by random with their origin replicate. The tools we are using for this analysis are chi-square test, histogram, possession process, hypothesis testing and goodness of fit test. In R we will find out the numbers of intervals and do the simulation test 2000 times.

Data

This time out data is based in 1990 the CMV sequence was 229,354 long and they would sceen all of the patterns. After screen, they found out that there are 296 long palindromes and those are at least 10 letters long. To be more pacific, the locations 12719, 75812, 90763 and 173893 have a 18 letters long sequence. In the data set, every single data point represents a location where the palindromes locate. All data points are valid.

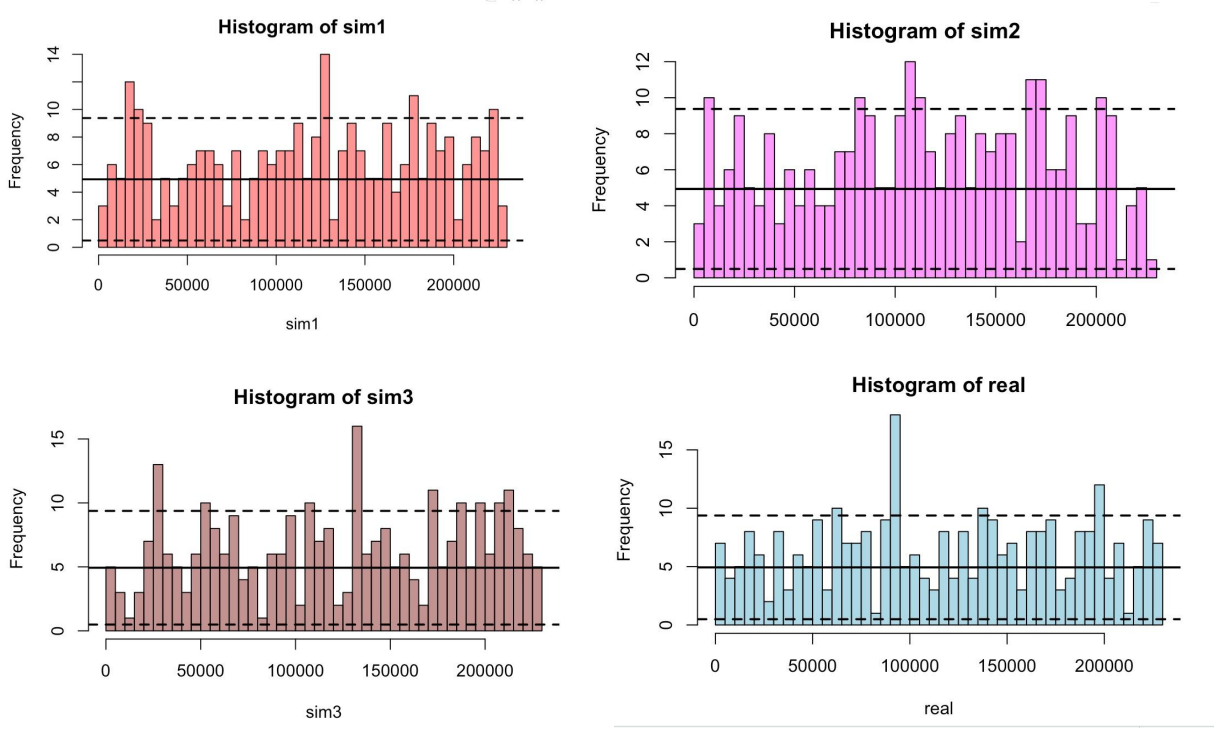
2.0 Basic Analysis

2.1 Random Scatter

Method

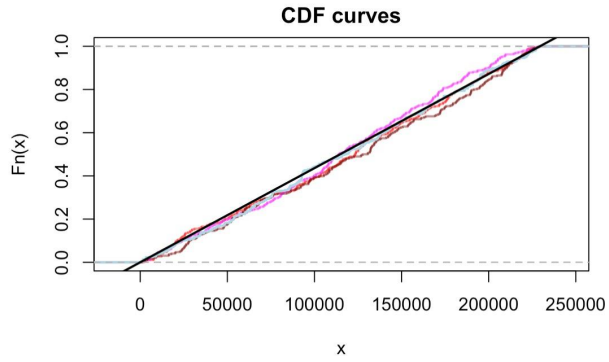
By using R, I will do the simulation study with the same sample size 294 and same total length 229,354. First do the random scatter plot to see how it looks like. Then by doing so several times, compare it with real palindrome locations. Graphically showing it in one graph and then analyzing the behavior.

Analysis



The above graphs are histograms for three simulations and real data. For each simulation, they are randomly selected 294 data points from 229354. And for the real data histogram, every 294 data points represent a location of the palindrome. Due to the sampling viability, there is a 95% variability bundle all datas included there are counting as generals. This bundle is represented as two dot lines. Apparently, all of the groups have some clusters of palindromes that surpass the upper dot line. For example, in the simulation one, there are clusters of palindromes around the 23000th, 125000th, and 156000th pairs of DNA. This condition also happened in simulation two and three. So it seems that even the randomly sampled data sets may appear in some clusters. Moreover, looking through these top bars, in these three simulations, the highest bar is around 15 or 16, which does not exceed 15 too much. However when the histogram of real show up, it is

clear to see that this histogram also has some cluster of palindromes such as 90000th and 190000th pairs of DNA. But the top bar here is far more than 15 which happened around 90000. While the shape of the histogram of real data is very similar to the others. Here we can conclude some similar patterns or feathers for these histograms. First, no matter simulations or our real data, they all contain some clusters or some bars that exceed the upper dot line. Second, the mean of these four data sets are all around 5 which are shown as black horizontal line in each histogram. Third, only based on these graphs, we may see that the number of data points in each interval are independent of each other. Therefore, roughly our real data meets the requirement of the uniform random scatter model. In other words, our real data seems randomly selected, uniformly distributed, but not really due to the abnormal behavior for the highest bar of histogram. Our real data's is higher than simulations. We also could plot their cumulative density function curves to check their distribution.



The black line in this group is the expectation line representing the standard CDF for uniform distribution. Other lines are the cdf for different data sets, the color here is matched the histogram shown above. Apparently, almost all lines are close to the expectation line in the graph, and it means that there are less errors on the histogram. We could conclude that palindrome locations are so far roughly uniformly distributed.

Conclusion

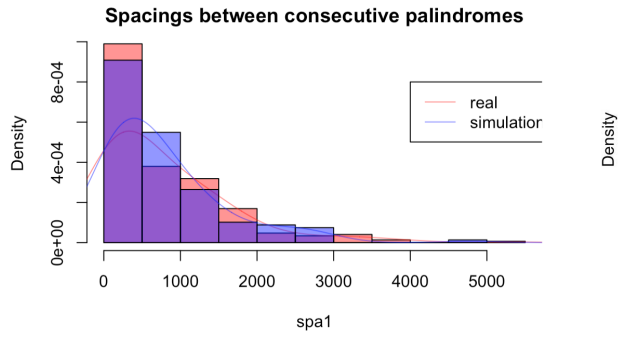
By doing the simulation study and comparing the palindrome location several times. The graph shows that the distribution is very similar to the real distribution and there are all clusters of palindromes in all the histogram graphs which means we cannot conclude that it is a uniform distribution just by looking at the histogram. Therefore, we use the CDF curves to see their distribution relate to the expected proportion of the expectation frequency. The line is really close to each other so we could conclude that the palindrome locations are uniformly distributed.Since this is a uniform distribution, then the cluster is not by chance.

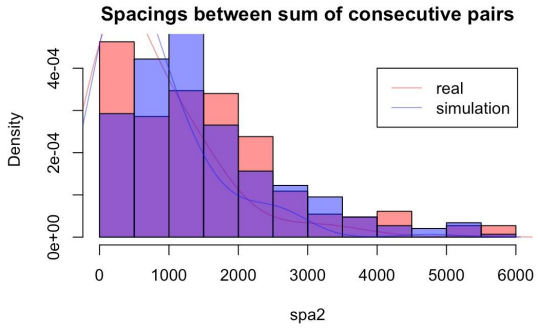
2.2 Locations and spacings

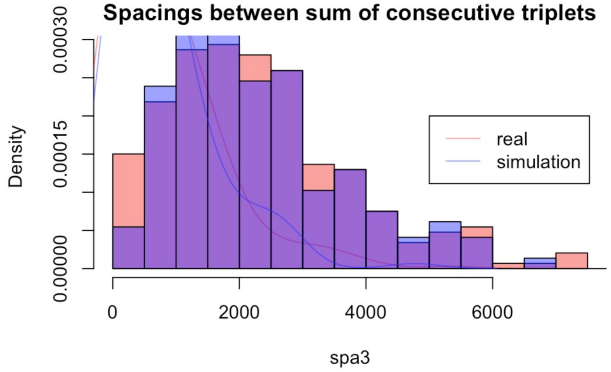
Method

In R, using the diff function and for loop to find the spacings between consecutive palindromes, and the sum of consecutive pairs, triplets. Then graphically showing them as histograms. Then plot the cumulative density function and check if they follow the exponential and gamma distribution.

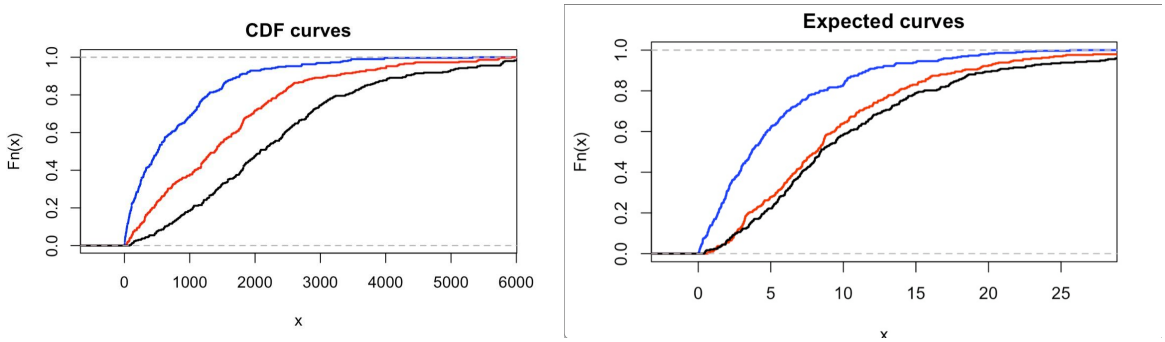
Analysis







In the histograms above, straightforwardly, spacing between consecutive palindromes are really similar to each other. From the graph, we could see that the color of purple represents the overlaps, the color of orange represents the real and the blud represents the simulation. Since the overlaps hold most of the spacing in the graph, the real and simulation are really similar to each other. Simulation is generated by random selection from 229,354 datas. Due to this fact and the huge overlapping areas for real data and simulation data, we could conclude that our real data is roughly uniform distribution. Same conclusion as prior analysis. Moreover the density lines also behave similarly for each other which emphasize the conclusion we get.



In the graph above, both the real CDF curve and expected exponential curve are blue. While the red line represents spacing between the sum of consecutive pairs and black is for triplets. After plotting the distribution into the CDF curve and the expected curves, we found something new. In general if our real data is randomly selected corresponding to uniform distribution, its single spacing should follow the exponential distribution. And the spacing between the sum of consecutive pairs and triplets should follow the gamma distribution. Comparing these two graphs, we can easily find the trend between these three curves that blue one is always on the top and red in the middle then is the black. Both expected and real curves follow this trend. But if we look more detailedly, we can find that the spacing between consecutive palindromes for our real data seems to follow the exponential distribution. While the gamma distribution seems not followed by pairs and triplets. Both expected gamma distribution lines for pairs and triplets should squeeze together. The gap is really small, almost overlapping each other. By seeing their frequency level, it shows that there is a large gap between the different sum of consecutive and spacing for real data. Moreover, the expected gamma distribution line for pairs is also lower than the real CDF curve in the first graph. However, the black line represents the triplets that seem to match each other. But still not very exactly. Overally, the trend is blue on top, red in the middle and the black in the bottom. By comparing it to the expected curves, we could see that the blue line is partially above the red line and the red line is roughly above the black line. Therefore, our data is not pretty uniformly distributed.

Conclusion

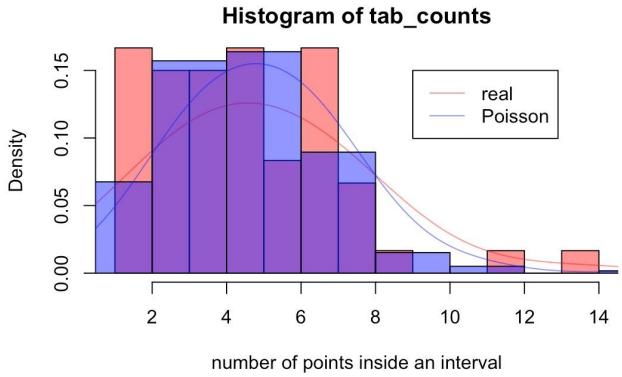
The graphs are showing that blue is located on the top and red is located in the middle and the last is black. They represent the one single spacing between consecutive palindromes, and sum of consecutive pairs, and the triplets. By simply looking through and analyzing the histograms of these different types of spacing, we can find that the overlapping areas are quite large, which may conclude the real data is probability uniform distribution. However after comparing the CDF curves and their corresponding distribution curves, exponential and gamma distribution respectively, we can only say that our real follows the trends that spacing of single points is higher than sum of pairs which is higher than sum of triplets. The real data is not exactly uniformly distributed due to the unexpected behavior of the cumulative density function curve.

2.3 Counts

Method

Graphically, make the histogram of counts in each interval. Then by the statistic test, chi-square test specifically, we will numerically check if they are uniformly distributed. Moreover, with the help of residuals, we could also make the standardized residual plot to check if they are uniform distribution.

Analysis



Above shows the histogram of counts for real data and its corresponding poisson distribution counts. Due to the theory that if we divide real data into intervals, its counts should follow the Poisson distribution. So here we made the graph to check that. Here the interval number is 60. The orange one is the distribution of counts of real data and blue one is the corresponding poisson distribution histogram. We could see the shape of the poisson distribution graph is not very really similar to the real distribution.The real data has higher density in 0~1 and 6~7 than poisson distribution.Even there is high overlap density, we can only say that the real data is somehow the uniform distribution, but not very exactly

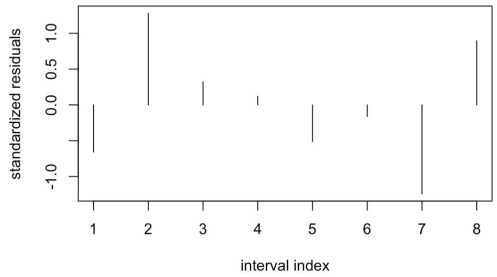
levels <fctr>	Observed <int>	Expected <dbl>
0	0	0.4321473
1	4	2.1319267
2	6	5.2587525
3	9	8.6477263
4	9	10.6655292
5	10	10.5233221
6	5	8.6525093
>=7	17	13.6880866

. 8 rows

```
        Chi-squared test for given probabilities with simulated p-value
        (based on 2000 replicates)

data:  real_trunc
X-squared = 4.8172, df = NA, p-value = 0.6702
```

By using R, we constructed a chart shown above which contains the observed number for each count and the expected values. By doing the chi-square test in R, we found that our test stats is 4.817, which induces a p-value of 0.6702. Even this p-value is far away from the significant level that we usually use 0.05, somehow this 0.6702 gives the reason for use to conclude our real data is uniformly distributed. However, this number is not high enough to state that our real data exactly fits the uniform distribution.



What's more, when we look at the standardized residuals graph which represents the differences between expected value and observed value. The longer the black line is the bigger difference they are. Corresponding to the conclusion we made by the chi-square test above, we can see that the standardized residuals are no more than 1.5. This means that the values of standardized residuals are really small and quite fit to our interval. Standing on an accurate side, we can still state the conclusion that the real data is uniform distribution, just likely.

Conclusion

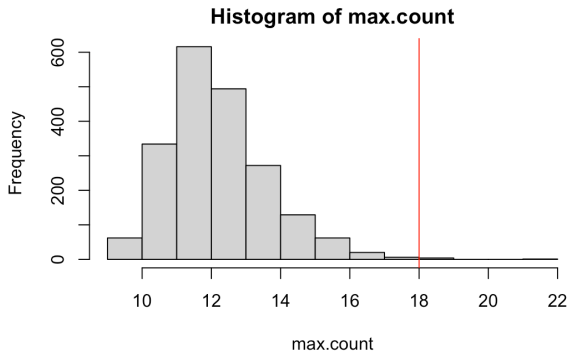
By doing the hypothesis test we see that the result actually fails to reject the null hypothesis which means that our real data is somehow uniform distribution. But due to the p-value here is only 0.6702 which is not big enough that we can only conclude that our real data is somehow uniformly distributed but not very exactly. Same conclusion derived by the standardized residual plot.This conclusion perfectly matches all conclusions we derived before.

2.4 The biggest cluster

Method

Find the highest counts in our real data set, then do the simulation study to check if this number is common in uniform distribution.

Analysis



So as to check if our max count in real data is common or not, we do the simulation study 2000 times. Every time we set the interval as 60 as we did in the very most and collect every max count in 2000 repeating experiments. Then we derived a histogram of max count shown above. The real vertical line in the histogram represents the max count in real data, which is 18. First, the number of intervals setted as 60 is pretty reasonable, since if the number of intervals is less which means the regions examined are large then the maximum count is large. For the opposite side, if the number of intervals is too much then the regions will be small. If the regions are small then the maximum count will be small. Thus 60 is both informative and reasonable. In the histogram shown above, we could easily see that 18 is quite abnormal among these 2000 repeating experiments. Even though there are indeed some outliers happening here which is 22, a number is even bigger than 18, the general max count for uniform distribution is between 10 to 14 under 60 intervals. And the probability of 18 is the max count is only 0.35%. Remark that we also discovered that in the very first analysis part, among three simulations, our real data indeed behaves abnormally. The highest bar in the histogram is much higher than others. Combining with the fact we found in the first analysis and here, we could confidently conclude that our real data is not exactly randomly selected, or uniformly distributed. There is an apparent big cluster occurring around 90000th location.

Conclusion

The interval with the greatest number of palindromes did indicate a potential origin of replication. When we did the simulation test, we figured out that, under the interval number as 60-- a really reasonable number, the common max. count is between 10 to 14. Our real data is not exactly randomly selected, or uniformly distributed. This means that the palindrome clusters occur by chance.

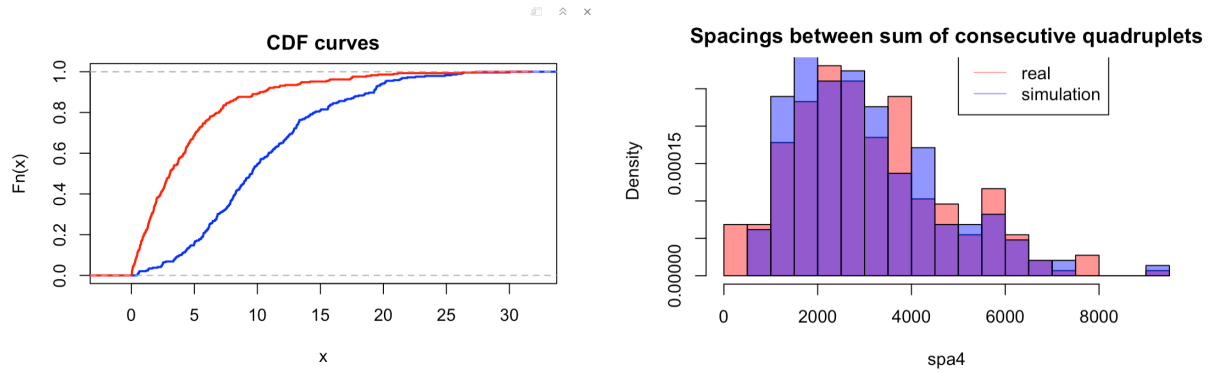
3.0 Advanced Analysis

After looking for the relationship between spacing between consecutive palindromes and sum of consecutive pairs, triplets. In this part, we will keep going along this direction but far more beyond. We will examine the spacing between sum of consecutive quatriclets.

Method

By using R, first find the spacing between consecutive quatriclets by for loop and converting them to a vector. Then plot the cdf curve of the spacing for both real data and the expected curve which is the gamma distribution curve. Also by the histogram, compare the spacing for real data and simulation data.

Analysis



The red curve in the first graph represents the expected cumulative density function(gamma distribution) curve, and the blue curve is derived from real data. Pretty similar to the conclusion derived from previous analysis. Our real data does not follow the expected distribution. If the real data is exactly uniformly distributed, its spacing between consecutive quatriclets should follow exactly the gamma distribution shown in the red line in the graph. But the blue curve is lower than the red curve. While, in the histogram of the spacings between sum of consecutive quadruplets, the overlapping area is quite large,which means the distribution of the quadruplets spacing behaves similarly. For real data there is some spacings are quite small which is around 0, the simulation' s is relatively concentrated showing as blue bars higher than overlaps. But both have some exceptions, some outlines.

Conclusion

There is a similar conclusion compared with the prior analysis, that our real data is not very randomly selected. Even the overlapping area with simulation data which is randomly selected is quite large, the spacing does not follow the gamma distribution which should follow if it is uniform distribution, or in other words random selected.

4.0 Conclusion and Discussion

After taking 229,354 letters long DNA sequence of CMV, we use the one with at least 10 letter long. There are 296 palindromes that fulfill this requirement. Those short letters will be ignored. Those 296 palindromes are located in the genome of the particular palindromes. After doing the analysis from several different aspects, we did find a cluster that was located around 90000th locations. These 296 palindromes behave pretty similarly to simulated datas. And we try to compare the spacing between consecutive points and the sum of consecutive pairs, triplets to its corresponding expected distribution. Mostly, the real data follows the trends of expected distributions but behaves not exactly the same. On advance analysis, we examine the spacing between sum of consecutive quatriclets, and derive the same conclusion. Then the chi-square test and standard residual plot also derive the result that these 296 locations somehow seem uniformly distributed but not very exactly. The p-value is only 0.6702 which is big enough in most of the cases but if we would like to study the DNA, this p-value would not be persuasive at all. Finally the 2000 times simulation also brings us the same conclusion that there is indeed a huge cluster.

These 296 palindromes are somehow uniformly distributed, but some abnormal big clustered exists. Mostly, in all the aspects we did the analysis here, the reason that there is always something abnormal happened is the existence of this huge cluster.

Based on the result we get, my advice to biologists who are about to start experimentally searching for the origin of replication that taking care and paying much attention to the region around 90000th place. There is a huge cluster which is not by chance, from where you may perhaps find the origin