

MATH 189 Project 3

Group 2, Winter 2021

Members:

- Kasen Teoh
- Chung En Pan
- Nathan Fallahi
- Parsa Ganjooi
- Eamon Jarrett-Mann

Introduction

The past year has brought the danger of viruses to the forefront of people's thoughts. Even if there are vaccines or treatments for certain viruses, mutations in a viruses' genetic code can render these less effective. The human CMV (Cytomegalovirus) is a life threatening disease to those who suffer from suppressed and deficient immune systems. In order to combat the virus, knowledge of how the virus replicates itself is extremely essential. Each time the virus replicates, a part of the original virus cell's DNA is replicated. Since the virus reproduces through replication, there are patterns and palindromes in the DNA which may hint at the site of replication and by identifying the site, scientists may be able to reduce or inhibit the virus from replicating. In order to find the origin of replication, we analyze the DNA, looking for any clusters and unusual clusters of palindromes. Upon the completion of this project, we would be able to inform scientists of the possible replication sites within the human CMV DNA.

Big question: Find the replication site of the herpes CMV virus.

1. **Simulate a random scatter** that corresponds to the locations of the palindromes on the DNA strand. Keep in mind the length and structure of our data. you will need to be able to generate a number of random scatters.
2. **Focus on locations and spacings.** Compare the observed distribution of location and spacings to the theoretical one coming from the random scatter. Use at least one simulation based test, and at least one graphical tool, and at least one formal hypothesis test. For locations consider and investigate at least consecutive pairs and triplets.
3. **Focus on counts.** Use graphical methods and more formal hypothesis tests to examine the counts of palindromes in the various regions of the DNA and compare the distribution of the observed to the one obtained from the random scatter. Split the DNA into non overlapping regions of equal length to do so. Vary the length choices (at least two choices) and perform sensitivity analysis (do the answers change when you change the length of the intervals; be cautious not to use too small or too big choices of length) . Consider at least one way to classify the regions according to their number of counts.
4. **Study the biggest cluster.** Does the interval with the greatest number of palindromes indicate a potential origin of replication? Perform a formal statistical test and propose at least one test of your own (graphical, simulation based, or formal). This last component is your independent question for Project 3.

Data

We used the file labeled hcmv.txt to look at the data. For the data we looked at palindromes that were at least 10 letters long with the longest one being 18 letters. We ignored anything that had less than 10 letters. The CMV DNA data that we looked at was 229353 letters long. Our data set has locations of 296 palindromes. The location can be anywhere from 1 to 229353 so the data is discrete. Our goal is to look at the data set to see if there are a lot of palindromes clustered close together. We are looking at the data given to us and we also need to simulate data in order to compare. To figure this out we'll be looking at 3 things which are location of palindromes, the distance between palindromes, and the number of palindromes per interval. These three things are used to identify a cluster, but for our question we are looking for the unusual clusters. The first step to organize the data would be to make intervals and within the interval we would count the number of palindromes in each interval. If you look at the data you can see at the 93000 and 195000 pairs of DNA you can notice a huge cluster of palindromes which does not depend on how large or small our intervals are. This shows us that these two locations are not due to chance. Next we would look at how many intervals there are with a certain number of palindromes in it. This is easy to see on a histogram because it is used as a density estimator. We use the same test statistic on both the observed data and the data we simulated.

Background

DNA is what makes every living thing different and unique. DNA is the code within every organism that contains the process of life and how to continue surviving. Within viruses, a part of the original virus cell's DNA is snipped off and then split open. Because DNA is composed of A, C, G, and T's, the virus now has two half strands of DNA. The virus' free nucleotides then join to these two half strands and now form two full DNA strands. The location of where the DNA is snipped off is essential to studying how to fight the virus and ultimately stop the virus from replicating and spreading throughout the immune system.

Without DNA, humans and other organisms would not exist. DNA is what gives people the specific traits that we experience everyday. Researchers are able to get conclusions purely off the gene sequence. With new technology it is also becoming easier to see the gene mapping and draw conclusions from that. This article does talk about how in 1996 when chip technology came out it made reading DNA a lot easier. The science behind all this is still growing today. On top of all this the article does also talk about how it is very rare today to see genes be cloned by digesting DNA restriction enzymes because of a new one step process that has been discovered (Leach, 1997).

CMV is of the Herpesvirus genus. For other members of this genus, such as the Herpes simplex virus (HSV), past research has shown connection to large clusters of DNA palindromes and origin of replication sequences (Lockshon & Galloway 1998). At the replication site, the DNA is split and forms two new DNA strands. Hence, there is a pattern at these sites, a palindrome in the DNA sequence. A palindrome is a sequence where it is the same forward and in reverse, for instance, ACGTTGCA. Therefore, in general, replication origins in DNA are considered major sites for regulating genome replication (Leung et al. 2005). However, a palindrome does not always represent a replication site. Due to the DNA sequence consisting of only four letters and the length of DNA, palindromes are bound to occur by random chance: statistically, palindromes are expected to occur half as often as non-palindromic sequences in double-stranded DNA (Lamprea-Burgunder et al., 2011). Consequently, we want to study the unusual occurrences of long palindromes to ensure we do not include the palindromes of random chance.

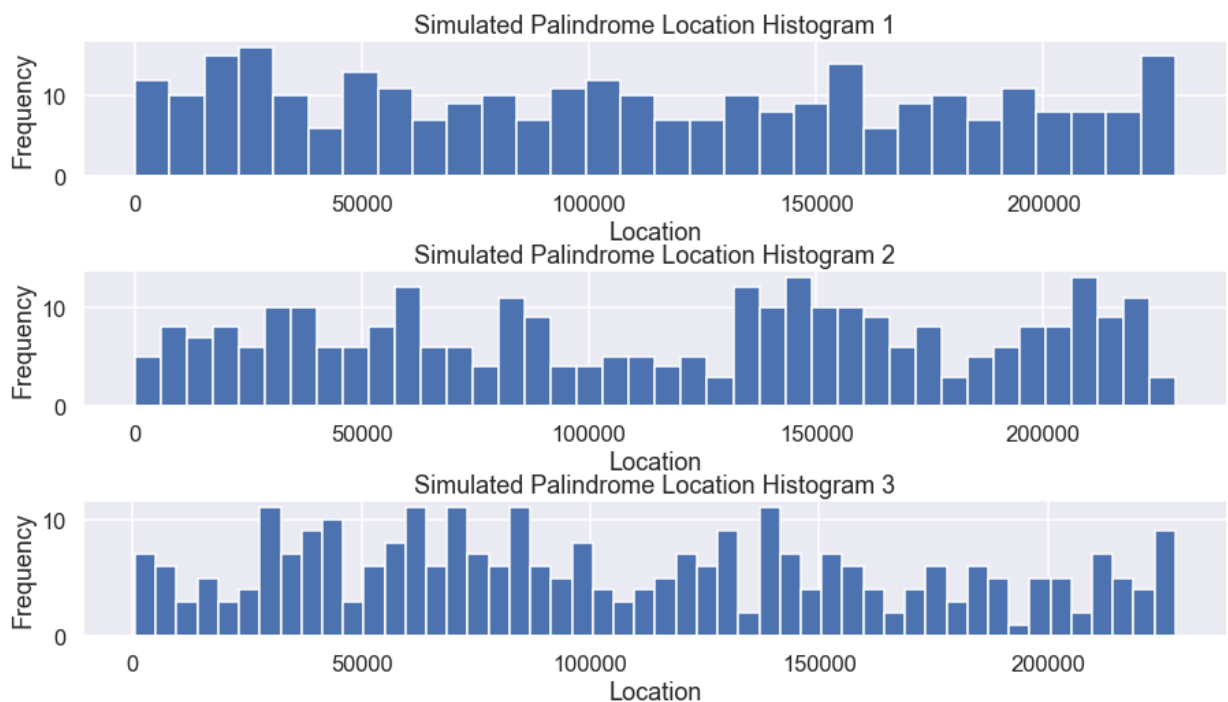
In particular, we want to study the Cytomegalovirus (CMV). Congenital CMV is a major cause of disability in children, with little evidence for change in disease burden over time in high- and middle-income countries despite large scientific and clinical advances in the CMV field (Manicklal et al., 2013). Most human CMV infections are asymptomatic or display minor symptoms. As a result, the CMV infection may go unnoticed in the human body and spread. While the CMV virus can lay dormant in a host, it becomes harmful when it enters the reproductive cycle, where it is able to replicate countless copies of itself. An untreated CMV infection can lead to a general infection with the involvement of viruses such as herpes and colitis and may be transmitted to the fetus and/or child (Schottstedt et al., 2010).

Investigation

Question 1

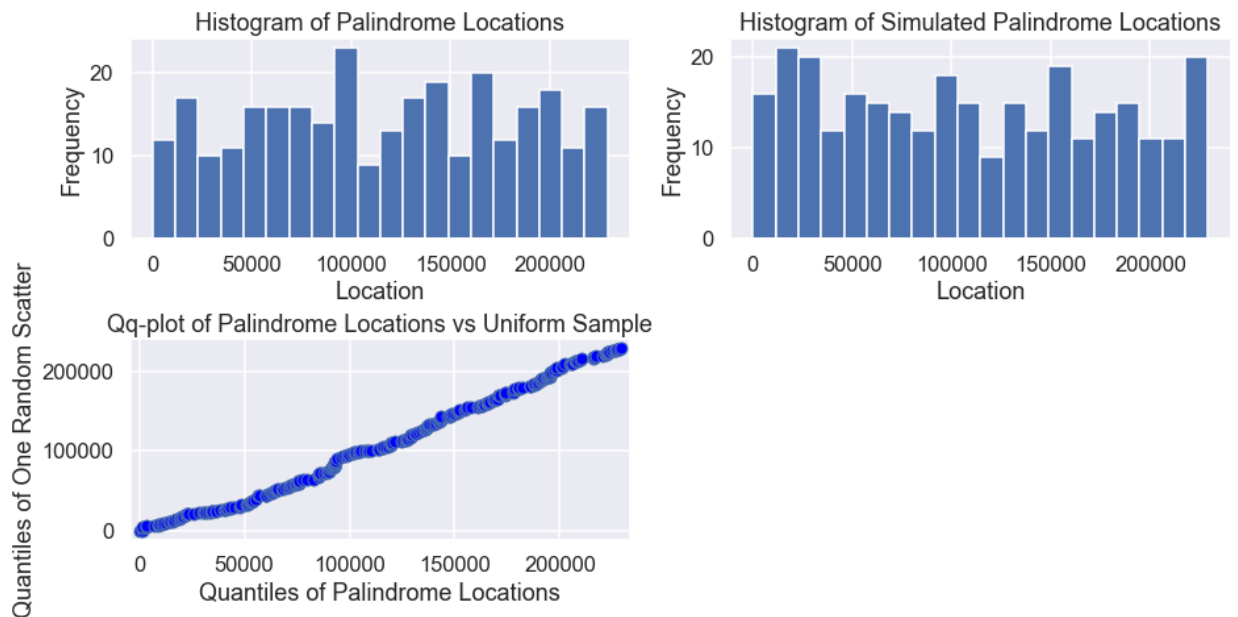
Simulate a random scatter that corresponds to the locations of the palindromes on the DNA strand. Keep in mind the length and structure of our data. you will need to be able to generate a number of random scatters.

Below, we see the histograms for 3 random scatters that correspond to the locations of the palindromes on the DNA strand. We see that distributions of palindrome locations for the simulated data are roughly the same. Hence, we choose one of the random samples to use as a comparison when carrying out our statistical and graphical tests.

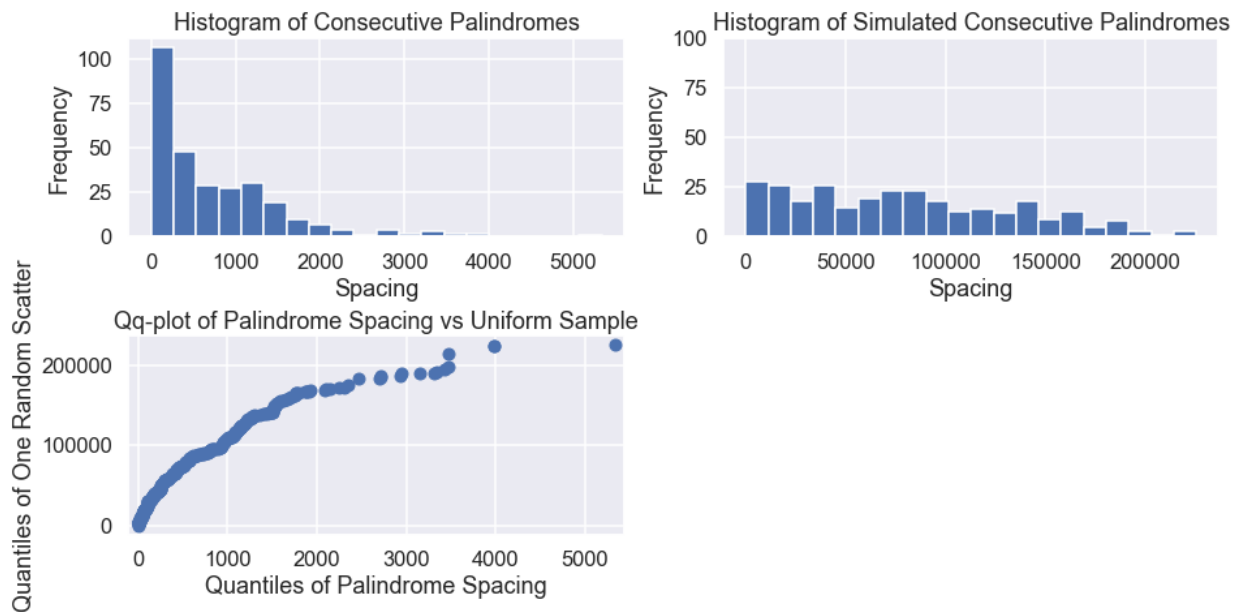


Question 2

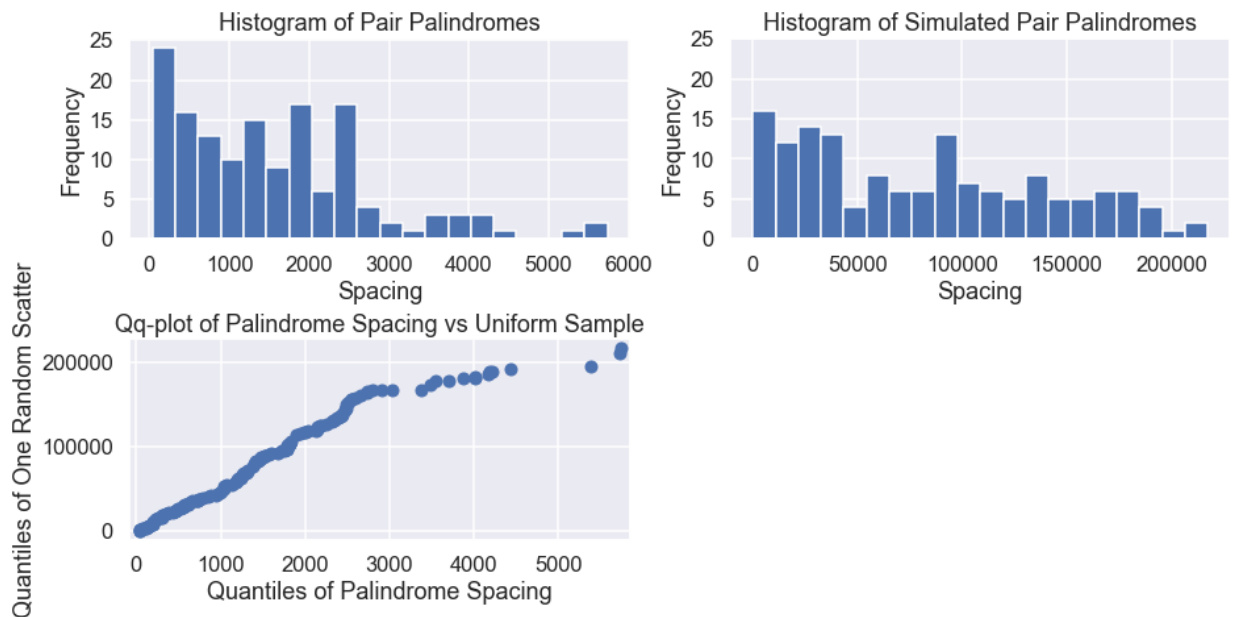
Focus on locations and spacings. Compare the observed distribution of location and spacings to the theoretical one coming from the random scatter. Use at least one simulation based test, and at least one graphical tool, and at least one formal hypothesis test. For locations consider and investigate at least consecutive pairs and triplets.



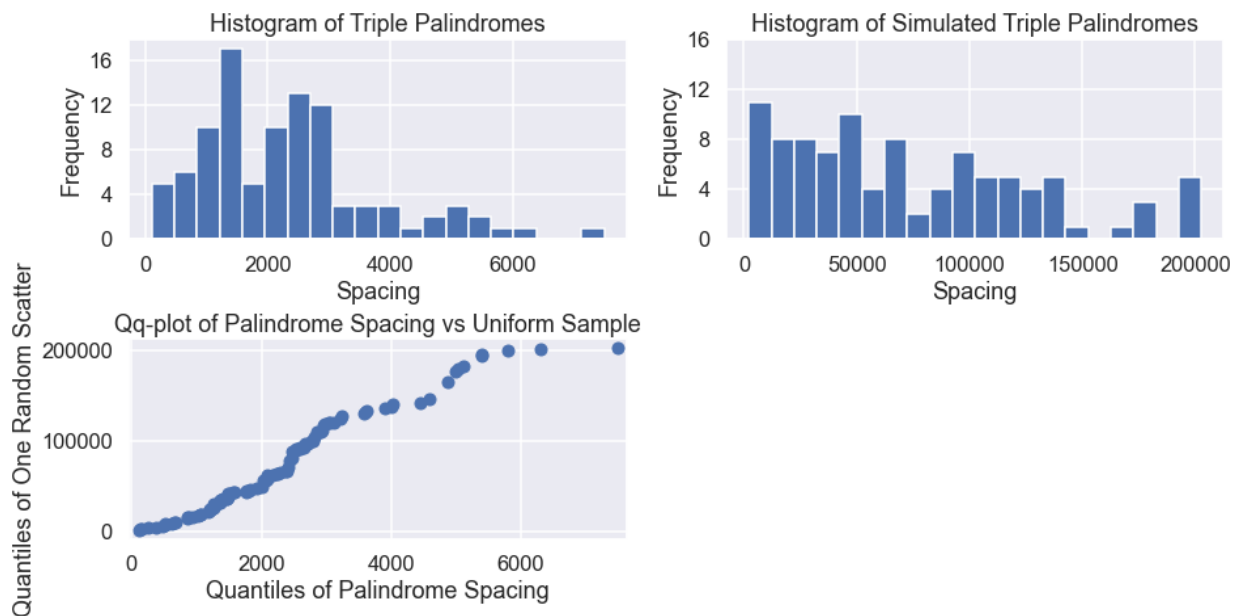
The three graphical distributions above show that the two samples are roughly the same in distribution which supports the thinking that the palindrome locations are random. However, from the two histograms, we see that the original data has a few peaks where there are more frequencies.



Visualizing the spacings between the original and simulated data, we see that in consecutive palindromes there are many more palindromes with only a few spaces between them. Additionally, in the quantile-quantile plot, we see that as the quantile grows, the spacing between the points becomes more evident. Hence, this leads us to believe that the two data sets do not have the same distribution.



Similarly to consecutive palindromes, we once again see that there is a large number of palindromes that has very few spacings in between them and less with larger spacings. In the simulated data, we see that there are a few pair of palindromes with ten thousands and even hundred thousands of spacings in between. Once again, the qq-plot is more spread out and further strengthening out hypothesis that the two data sets are not of the same distribution.



The above histograms all show that there are constantly more palindromes in the original data that have smaller spaces in between each other while the simulated data has larger spacings. As for the qq-plots, aside from the location of the palindromes, we see that when plotting the spacings between the palindromes against the spacings in the simulated data, we arrive at the hypothesis that the two do not follow the same distribution.

Comparing the two data sets, there are many palindromes with small spacings in between them in the original data set, which is to be expected because our alphabet only consists of four letters (A, T, C, and G).

Location of Palindromes

H_0 : The location of the palindromes follow a uniform distribution

H_A : The location of the palindromes do not follow a uniform distribution

Consecutive Palindromes

H_0 : The spacing of consecutive palindromes follow an exponential distribution

H_A : The spacing of consecutive palindromes do not follow an exponential distribution

Consecutive Pair Palindromes

H_0 : The spacing of consecutive pairs palindromes follow a gamma distribution

H_A : The spacing of consecutive pairs palindromes do not follow a gamma distribution

Consecutive Triple Palindromes

H_0 : The spacing of consecutive triple palindromes follow a gamma distribution

H_A : The spacing of consecutive triple palindromes do not follow a gamma distribution

Hypothesis tests for original data

Out[9]:

	Parameter	Test Statistic	95% Quantile	Degrees of Freedom	P-Value	Alpha
Location of Palindrome	# of Palindromes = 14.8	18.054	18.624	19	5.19e-01	0.050
Distance of Consecutive Palindromes	Lambda = 1.29e-03	102.210	65.171	48	1.29e-05	0.050
Distance of Consecutive Pair Palindromes	Lambda = 6.44e-04	112.176	65.171	48	7.35e-07	0.050
Distance of Consecutive Triple Palindromes	Lambda = 4.29e-04	146.304	65.171	48	1.27e-11	0.050

In the cell above, we see that we fail to reject the null when looking at the location of palindromes as the test statistic is less than that of the 95% chi-square quartile and the p-value is much greater than that of the alpha. This means that we have insufficient evidence to say that the location of the palindromes do not follow a uniform distribution, i.e the location of the palindromes in our data follow a uniform distribution.

For the last three hypothesis tests, the three test statistics are greater than the 95% quantile and the p-value is less than that of the alpha. Hence, we are able to reject the null hypothesis in favor of the alternative. In other words, for consecutive palindromes, we have sufficient evidence to say that they do not follow an exponential distribution and for consecutive, pair, and triple palindromes, they do not follow a gamma distribution.

Location of Palindromes

H_0 : The location of the simulated palindromes follow a uniform distribution

H_A : The location of the simulated palindromes do not follow a uniform distribution

Consecutive Palindromes

H_0 : The distance of the consecutive simulated palindromes follow an exponential distribution

H_A : The distance of the consecutive simulated palindromes do not follow an exponential distribution

Consecutive Pair Palindromes

H_0 : The distance of the consecutive simulated pair palindromes follow a gamma distribution

H_A : The distance of the consecutive simulated pair palindromes do not follow a gamma distribution

Consecutive Triple Palindromes

H_0 : The distance of the consecutive simulated triple palindromes follow a gamma distribution

H_A : The distance of the consecutive simulated triple palindromes do not follow a gamma distribution

Hypothesis tests for simulated data

Out[10]:

	Parameter	Test Statistic	95% Quantile	Degrees of Freedom	P-Value	Alpha
Location of Palindrome	# of Palindromes = 14.8	22.108	15.024	19	2.79e-01	0.050
Distance of Simulated Consecutive Palindromes	Lambda = 1.23e-05	109.580	65.171	48	1.58e-06	0.050
Distance of Simulated Consecutive Pair Palindromes	Lambda = 1.22e-05	99.412	65.171	48	2.78e-05	0.050
Distance of Simulated Consecutive Triple Palindromes	Lambda = 1.34e-05	140.000	65.171	48	1.08e-10	0.050

In the cell above, we achieve the same results as the four hypothesis tests for the original data. We fail to reject the hypothesis test involving the location of palindromes, hence, it follows a uniform distribution.

On the other hand, the other three hypothesis tests, consecutive, consecutive pairs, and consecutive triples, we reject the null and so the three distance distributions do not follow an exponential, gamma, or gamma distribution, respectively.

Question 3

Focus on counts. Use graphical methods and more formal hypothesis tests to examine the counts of palindromes in the various regions of the DNA and compare the distribution of the observed to the one obtained from the random scatter. Split the DNA into non overlapping regions of equal length to do so. Vary the length choices (at least two choices) and perform sensitivity analysis (do the answers change when you change the length of the intervals; be cautious not to use too small or too big choices of length). Consider at least one way to classify the regions according to their number of counts.

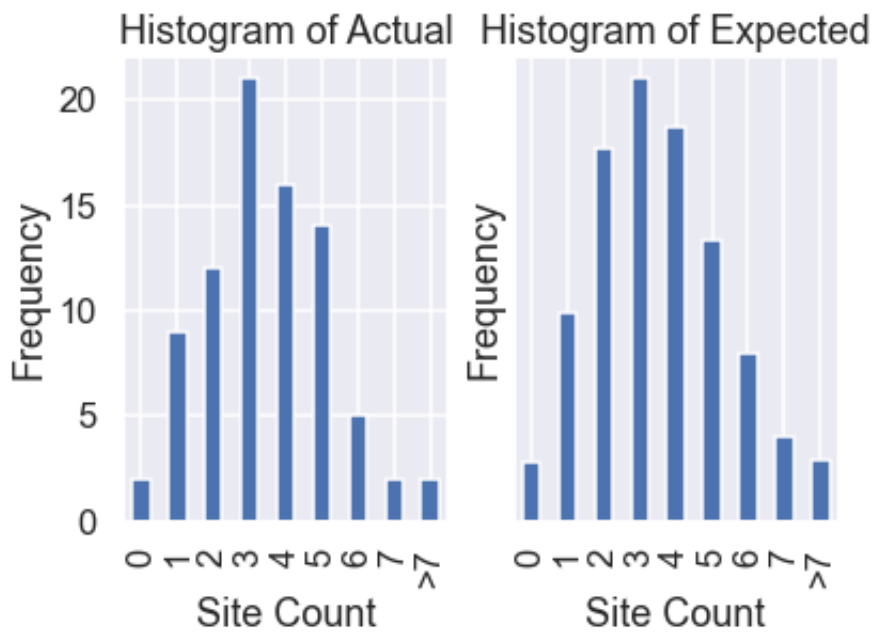
H_0 : the number of palindromes in a given interval are $\sim \text{poisson}(\lambda)$

Where the maximum likelihood estimator $\hat{\lambda} = \frac{\text{length of data}}{\text{length of intervals}}$

H_1 : This is not the case (there is more clustering than one would expect from randomness alone)

```
Test Statistic: 3.029
P-value: 0.933
for intervals of length: 2797
Fail to reject the null hypothesis
Classification by count
```

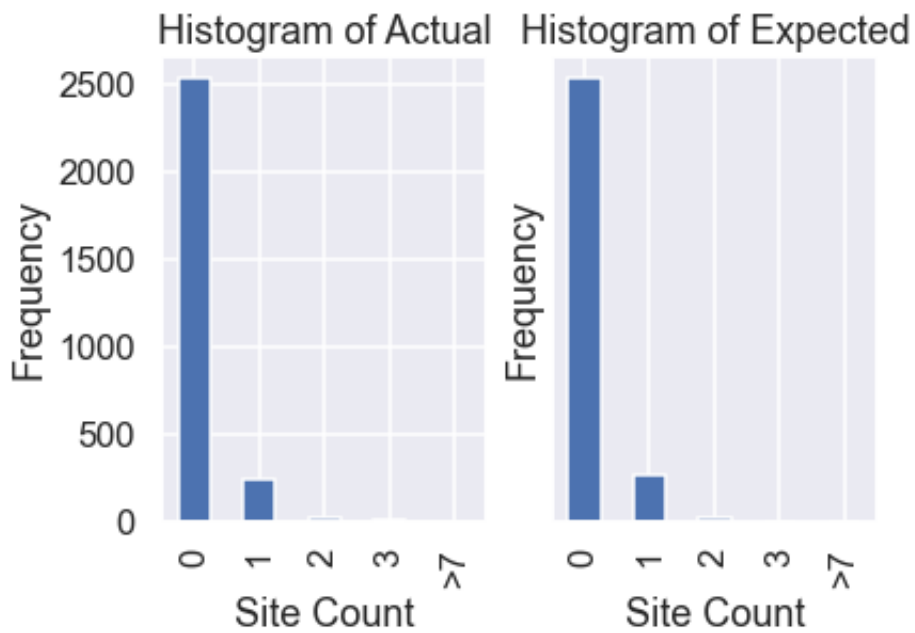
```
Out[17]: 0      improbable
         1      improbable
         2      improbable
         3      improbable
         4      improbable
         5      possible
         6      likely
         7      likely
         >7    highly likely
dtype: object
```



As you can see from the graph above, the occurrence of site counts is not too different from what one would expect from Poisson of λ .

```
Test Statistic: 13.824
P-value: 0.008
for intervals of length: 82
Reject the null hypothesis
```

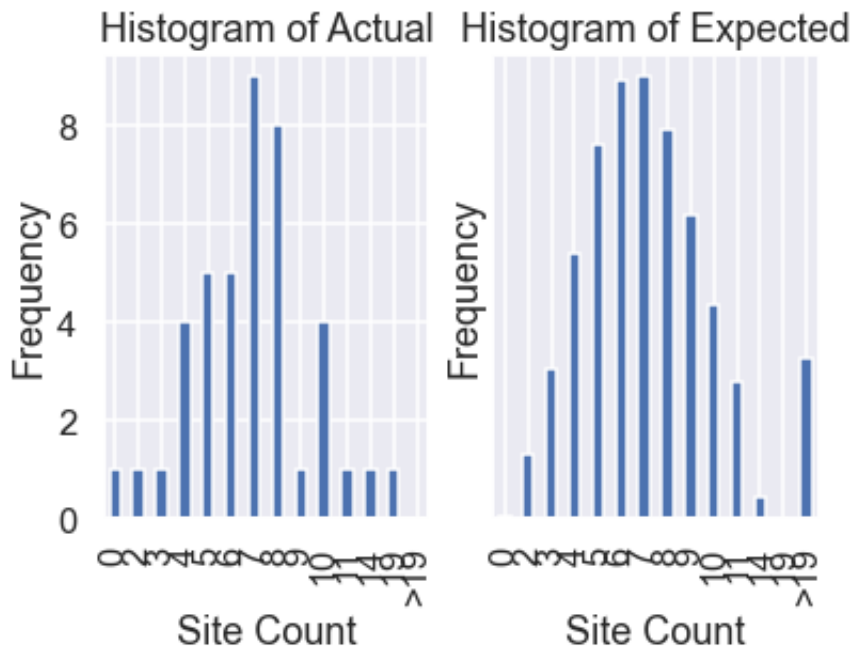
```
Out[19]: 0           possible
          1      highly likely
          2      highly likely
          3      highly likely
          >7     highly likely
dtype: object
```



As you can see from the graph above, the occurrence of site counts is not too different from what one would expect from poisson of lambda hat.

```
Test Statistic: 290.637
P-value: 0.0
for intervals of length: 5594
Reject the null hypothesis
```

```
Out[21]: 0      improbable
          2      improbable
          3      improbable
          4      improbable
          5      improbable
          6      improbable
          7      improbable
          8      improbable
          9      possible
         10      possible
         11      likely
         14      highly likely
         19      highly likely
         >19     highly likely
dtype: object
```

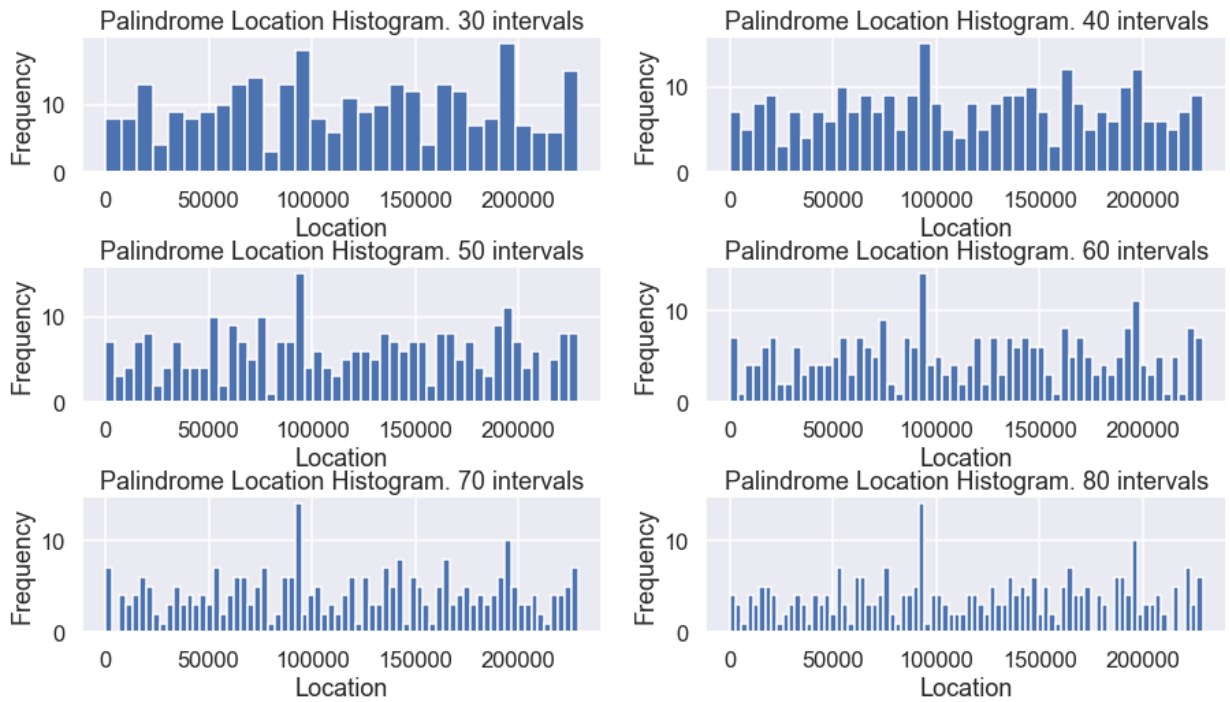


As you can see from the graph above, the occurrence of site counts is not too different from what one would expect from poisson of lambda hat.

Question 4

Study the biggest cluster. Does the interval with the greatest number of palindromes indicate a potential origin of replication? Perform a formal statistical test and propose at least one test of your own (graphical, simulation based, or formal). This last component is your independent question for Project 3.

In the following subplot of histograms, we can see a visual representation of the locations of palindromes from both our data. From what we can observe (and what was revealed to us in lecture), there are two prominent clusters of palindromes in our real data at roughly the 93,000th and 195,000th pair. Notably, the cluster at around the 93,000th pair is the biggest (only at 30 intervals or fewer is a different cluster bigger).



We can use a **scan statistic** to calculate the probability of seeing a maximum number of k palindromes within a single interval given m intervals. Let T_m be the maximum number of palindromes over m intervals. Let's create the following hypothesis:

- $H_0: T_m$ has a distribution of a maximum of independent Poisson(λ) random variables.
- $H_1: T_m$ does not have a distribution of a maximum of independent Poisson(λ) random variables.

In the below table we can see the calculated probabilities for T_{30} , T_{40} , T_{50} , T_{60} , T_{70} , and T_{80} . With an appropriately large number of intervals, it is very clear that there is an unusually large cluster of palindromes at around the 93,000th base pair which indicates a potential origin of replication.

Out[24]:

	lambda	Interval Width	k	Interval Location	p-value	Result
m						
30	9.867	7645	19	(190823.667, 198449.533]	0.172	Fail to reject the null hypothesis
40	7.400	5733	15	(91687.4, 97406.8]	0.308	Fail to reject the null hypothesis
50	5.920	4587	15	(91687.4, 96262.92]	0.060	Fail to reject the null hypothesis
60	4.933	3822	14	(91687.4, 95500.333]	0.036	Reject the null hypothesis
70	4.229	3276	14	(91687.4, 94955.629]	0.009	Reject the null hypothesis
80	3.700	2866	14	(91687.4, 94547.1]	0.003	Reject the null hypothesis

Additionally, we can compare our data to random samples from a Poisson distribution. Again, let k be the maximum number of palindromes in any interval and let m be the number of intervals. We draw 10,000 random samples from a Poisson distribution given our computed λ . Then, we determine the 99th percentile for the maximum number of palindromes in a given interval from our random sample and compare it to the value we observe in our real data. Our hypotheses are thus as follows:

- H_0 : The maximum number of palindromes in an interval follows a Poisson distribution.
- H_1 : The maximum number of palindromes in an interval does not follow a Poisson distribution.

In the following table we conduct this test for 30, 40, 50, 60, 70, and 80 intervals. For each, we can see that our observed maximum number of palindromes in an interval is larger than the 99th percentile from a Poisson distribution. Therefore it is clear that the largest cluster of palindromes in our data is unusual.

Out[25]:

	lambda	Interval Width	k	99%	Interval Location	Result
m						
30	9.867	7645	19	18	(190823.667, 198449.533]	Reject the null hypothesis
40	7.400	5733	15	14	(91687.4, 97406.8]	Reject the null hypothesis
50	5.920	4587	15	12	(91687.4, 96262.92]	Reject the null hypothesis
60	4.933	3822	14	11	(91687.4, 95500.333]	Reject the null hypothesis
70	4.229	3276	14	10	(91687.4, 94955.629]	Reject the null hypothesis
80	3.700	2866	14	9	(91687.4, 94547.1]	Reject the null hypothesis

Conclusion

Following the chi-square tests for a specified distribution, we arrived at the conclusion that the location of the palindromes does follow a uniform distribution. However, the spacings of consecutive, non-overlapping pairs and triple palindromes do not follow an exponential distribution. We had also found that for intervals of length 82, 2797, and 5594, that we do not have sufficient evidence of a poisson distribution for length 2797. For lengths 82, and 5594, we had rejected the null hypothesis in favor of the data not having a Poisson distribution.

The evidence from the hypothesis testing of the distribution of counts overall seems to indicate that the distribution could plausibly be a poisson of the maximum likelihood estimator. However, given the results of the scan test and our invented test, it seems likely that the data is not the result of uniform chance, as the highest palindrome count cluster seems to be exceptionally unusual.

From this, our analysis finds that there is an unusually large cluster of palindromes at the 93,000th base pair that is a potential origin of replication.

Theory

Poisson process

The Poisson process is a model for random occurrences such as waiting time for a bus, arrival time for customers waiting in the queue. This process is distributed with no regularity and follow with the characteristic that:

- rate (λ) at each hit does not change with each location
- the hit in each region are independent
- one hit can only be in exactly one location

Detecting the unusual scatter of the palindromes, the poisson process becomes very useful as the process is a model for uniform random scatter, which represents a good reference model for the DNA strand, as DNA strands are composed of A,C,T, and G. DNA strand can be treated a line that the palindromes are randomly scatter across the line, the location of each palindromes are unique with no overlapped region and the hits of the palindromes in certain interval are independent to other interval. These all align with characteristics for the poisson process. With the properties of the Poisson process, we can check whether Poisson process is a reference model for the DNA strand data.

Chi-Square Test

A chi-squared test is used to determine whether a data comes from a specific distribution in categorical variables. This is used to compare the observed frequency versus the expected frequency to check whether the data is statistically significant. The null hypothesis is that the data does follow the specific distribution while the alternative hypothesis is that the data does not follow the specific distribution. We are able reject the null hypothesis if the observed value of the test statistic is larger than the 95% quantile or is the p-value of the test-statistic is greater than the specified alpha level. The data distributions that were tested against the data were uniform distributions, poisson distributions, exponential, and gamma distributions. The formula for the test statistic is

$$\sum_{i=1}^m \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

We used the chi square test in question two to compare the observed location and spacing to the theoretical or expected location. When comparing the test statistic, we found the 95 percentile of the chi-square distribution using a degree of freedom of the number of bins minus the number of parameters estimated minus one $df = m - k - 1$ where m is the number of bins and k is the number of estimated parameters.

Poisson Distribution

The poisson distribution is used because it closely resembles the data that we are given to work with. The poisson distribution quantifies the counts of hits in fixed intervals of the poisson point process. That is, it is a way of estimating what the occurrence count of an event will be in some given interval. The poisson distribution's probability mass function is

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

We used chi-square testing to determine how well the observed palindrome counts within arbitrarily chosen intervals can be modeled by using a Poisson distribution. These are used to determine whether there are unusual clusters of palindromes, relative to what one would expect if it was determined by random chance alone. We chose to use a wide variety of intervals to ensure the cluster findings are not some strange result of the intervals chosen.

λ can be estimated by the sample mean \bar{x} .

Uniform Distribution

Analyzing the locations of the palindromes, we had split the data up into equal length intervals. Then, because the number of palindromes in each interval is known, we are able to test the data against a uniform distribution. Hence, the expected counts of each interval is the total number of palindromes divided by the number of intervals: Expected = $\frac{\text{total number of palindromes}}{\text{the number of intervals}}$, and because we did not need to estimate any parameters, the degree of freedom is the number of intervals minus one.

Exponential and Gamma Distribution

When analyzing the spacings between consecutive, consecutive pairs, and consecutive triple palindromes, we use the chi-square test to test again an exponential and gamma distribution because we are analyzing the distance until the next palindrome. In other words, we are measuring the time until the next hit, where the time is the distance and the hit is the location of a palindrome.

$$P(\text{no palindromes in an interval of length } t) = e^{-\lambda \cdot t}$$

When testing against exponential and gamma distribution, we estimated the parameter of λ , hence, the degree of freedom is the number of bins minus two. For exponential and gamma distribution, we had used $\frac{1}{\bar{x}}$ as an estimator for λ .

Scan Statistic

The scan statistic is used to find the unusual cluster. An example of an unusual cluster of palindromes is when a large number of palindromes are really close together. Before finding the unusual cluster we have to decide what an usual cluster will look like first. To do this we have to decide on the maximum number of palindromes per different intervals. It is important to choose the correct length so we are able to identify the palindromes and the number of them in each cluster.

The first step in this process is to see what the distribution of T_m is under the null hypothesis. In our case, we want to calculate the probability that the maximum number of palindromes in an interval, given m intervals, is greater than or equal to the max value we observe in our data k . This can be done by:

$$\text{p-value} = 1 - \left(e^{-\hat{\lambda}} \left[1 + \hat{\lambda} + \frac{\hat{\lambda}^2}{2!} + \dots + \frac{\hat{\lambda}^{k-1}}{(k-1)!} \right] \right)^m$$

Histograms

- Histograms are estimators of the probability density function for a given distribution. Histograms consist

of bins that count the number of observations within each interval. The number of bins must be predetermined.

$$\hat{f}_x(a) = \frac{1}{n}(\# \text{ of observations within } B_l) \times \left(\frac{1}{\text{length of bin } B_l} \right)$$

Where a is a specific location in the distribution and B_l is a bin that contains a

- Histograms have lower bias the more number of bins are used.
- Histograms provide useful information about the spread of the data and can be used to identify potential outliers, whether the underlying distribution is unimodal/bimodal or symmetric/asymmetric, and what proportion of the distribution lies above or below a certain value.

Contributions

- Kasen Teoh
 - I contributed to the introduction and background sections. I also helped with completing question 2 and writing the theory section for the exponential and uniform distributions.
- Chung En Pan
 - I contributed to the theory section. Focusing on the Poisson process, defining the characteristic of the Poisson process. I also helped with completing question 2.
- Nathan Fallahi
 - I contributed to the Background section and completed Question 4. I also added to the Theory section, specifically the information about the scan statistic and histograms.
- Parsa Ganjooi
 - I did some of the background section, did the data section, put some information in the theory section, and helped with the questions. In the theory section I added information to the chi-squared test and the scan test statistic.
- Eamon Jarrett-Mann
 - I completed question 3, wrote the theory section on the Poisson distribution/chi-square goodness of fit testing for that section & helped write the conclusion.

References

- Leung, M. Y., Choi, K. P., Xia, A., & Chen, L. H. (2005). Nonrandom clusters of palindromes in herpesvirus genomes. *Journal of computational biology : a journal of computational molecular cell biology*, 12(3), 331–354. <https://doi.org/10.1089/cmb.2005.12.331> (<https://doi.org/10.1089/cmb.2005.12.331>)
- Leach, D. R., Okely, E. A., & Pinder, D. J. (1997). Repair by recombination of DNA containing a palindromic sequence. *Molecular microbiology*, 26(3), 597-606.
- Lamprea-Burgunder, E., Ludin, P., & Mäser, P. (2011). Species-specific typing of DNA based on palindrome frequency patterns. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 18(2), 117–124. <https://doi.org/10.1093/dnares/dsr004> (<https://doi.org/10.1093/dnares/dsr004>)
- Lockshon, D., & Galloway, D. A. (1988). Sequence and structural requirements of a herpes simplex viral DNA replication origin. *Molecular and Cellular Biology*, 8(10), 4018-4027.
- Manicklal, S., Emery, V. C., Lazzarotto, T., Boppana, S. B., & Gupta, R. K. (2013). The "silent" global burden of congenital cytomegalovirus. *Clinical microbiology reviews*, 26(1), 86–102. <https://doi.org/10.1128/CMR.00062-12> (<https://doi.org/10.1128/CMR.00062-12>)
- Schottstedt, V., Blümel, J., Burger, R., Drosten, C., Gröner, A., Gürtler, L., Heiden, M., Hildebrandt, M., Jansen, B., Montag-Lessing, T., Offergeld, R., Pauli, G., Seitz, R., Schlenkrich, U., Strobel, J., Willkommen, H., & von König, C. H. (2010). Human Cytomegalovirus (HCMV) - Revised. *Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft für Transfusionsmedizin und Immunhämatologie*, 37(6), 365–375. <https://doi.org/10.1159/000322141> (<https://doi.org/10.1159/000322141>)