Case Study: Anomaly Observation in Palindromic Sequences

MATH 189: Exploratory Data Analysis and Inference, Project 3

Alison Camille Dunning [1], Bryan Talavera [2], Jared Dishman [3], Abdiaziz Weheliye [4], Yaoxin Li [5], Ruotian Gao [6], Ye Yint Win [7]

[1] Second Year Undergraduate: Data Science, A16166698
[2] Second Year Undergraduate: Data Science, A13278593
[3] Transfer Undergraduate: Data Science, A16688060
[4] Transfer Undergraduate: Cognitive Science and Machine Learning, A15690733
[5] Transfer Undergraduate: Data Science, A16542726
[6] Fourth Year Undergraduate: Data Science, A15230548
[7] Transfer Undergraduate: Data Science, A16688105

Introduction

In this paper, we bring into focus viral spread through sources of replication in DNA sequences. Our motivation is the effects of the cytomegalovirus (CMV), a life-threatening disease that poses the greatest risk to hosts with deficient immune systems. To combat this virus we are trying to locate where DNA replicates which is called origin of replication, this is where the DNA contains instructions for reproduction for that virus. If we understand how the DNA virus replicates we can reverse the process and the replication process would never take place, to find the origin of replication, a DNA is cut into segments and each segment is tested to see whether it can replicate. If it cannot replicate then the origin of replication must not be contained in that segment. A DNA sequence is a coded message that consists of only a four letter alphabet, ACTG, since there are only four letters the DNA sequence would have many patterns. Complementary palindrome is one of those patterns in DNA, the letter G is complementary to C, and the letter T is complementary to A, complementary palindrome is a sequence of letters that read the same backward as the complement of the forward sequence. Our goal in this paper is to find unusual clusters of complementary palindromes.

Background

Found by Oswald Avery, Colin MacLeod and Maclyn McCarthy, Deoxyribonucleic Acid which is more commonly known by the general public as DNA is the substance that contains the information by which an organism regenerates its cells and serves as a carrier of genetic information and traits that every parent passes onto their biological offspring.[2] And in 1953, Rosalind Franklin, James Watson and Francis Crick found that DNA molecule has double helix structure consisting of three constituents: sugar (ribose), phosphate (phosphorous surrounded by oxygen responsible for its acidity), and nucleotide base

(adenine(A), cytosine(C), thymine(T), and guanine(G))[3]. Due to its nature, DNA replication is the process by which DNA makes a copy of itself during cell division. The process is carried out by an enzyme called helicase (usually proteins) which breaks the hydrogen bonds holding the complementary bases of DNA together (A with T, C with G).[4] Breaking the bond causes the separation of DNA, unzipping (snipping) helix into two single strands (Y-shape) which later combines with free nucleotides, creating new strands of DNA.

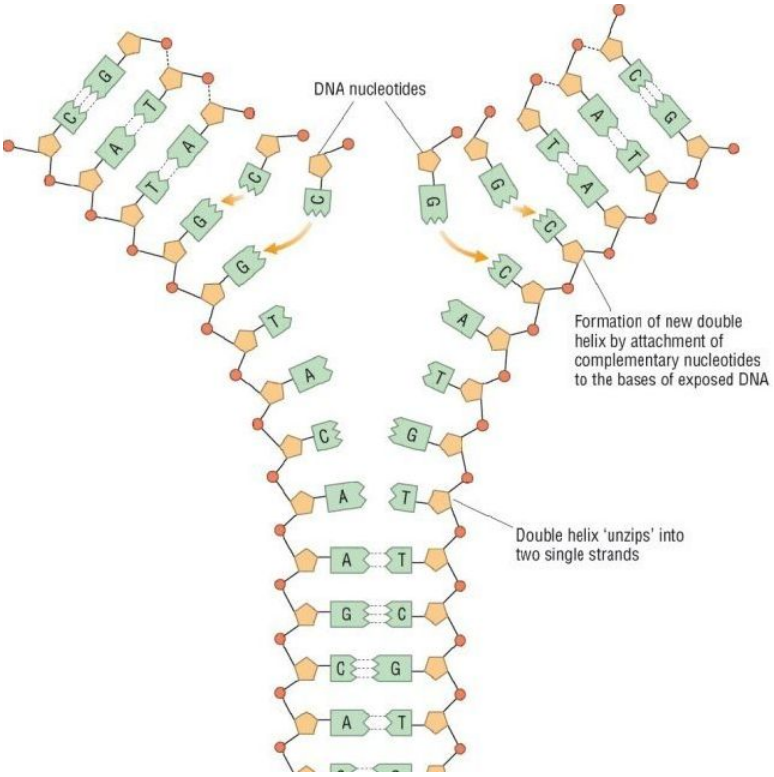


Figure 1. Process of DNA replication

Cytomegalovirus commonly known as CMV is a common virus which nearly one in three children are already infected by age five and over half of adults in the U.S. have in their body by age 40. It is related to the viruses that cause chickenpox, herpes simplex and mononucleosis or hepatitis, and it poses major risk for people in immunosuppressed states such as transplant patients and AIDS/HIV patients.[5] It can be spread through body fluids and once infected, CMV lays dormant and only becomes harmful when the virus enters a productive cycle in which it quickly replicated into tens of thousands of copies. The DNA for viruses typically ranges up to several hundred thousand base pairs in length.

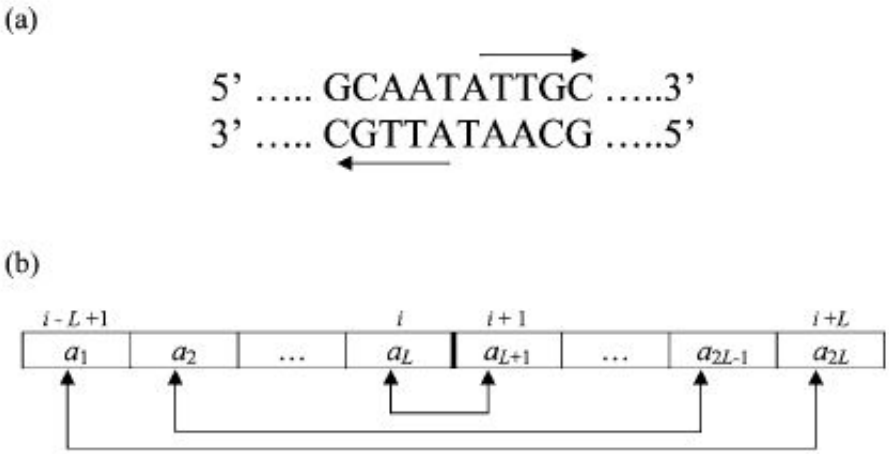


FIG. 1. DNA palindrome. (a) This is a palindromic nucleotide sequence on the two complementary strands of DNA which are always read in opposite directions from the 5′ end to the 3′ end as shown by the arrows. The displayed segment reads exactly the same on both strands. (b) On each strand, the first base of the palindrome is complementary to the last, the second to the second last, and so on. This is a schematic representation of such complementary pairing between the bases in a $2L$-palindrome centered at base i.

Figure 2.

And in order to study the CMV and how to combat it, we need to take a deeper look into usual clusters of complementary palindromes in different intervals of the DNA sequence. "Palindromes are symmetrical words of DNA in the sense that read exactly the same as their reverse complementary sequences" (Pg. 331)[6] By doing the scan test and identifying the intervals with unusual high concentration of palindromes, we can associate them with the replication origins on a few herpesviruses, one of CMVs, according to previous studies. Locating the origin of replication for CMV may help virologists find an effective vaccine against the virus which is why we will be at the data set to determine accurate prediction of replication origins.

Data

The data for this project is the DNA sequence of CMV published in 1990, and the data illustrates the palindromes in the DNA sequence. To be specific, there are 296 palindromes found at least 10 letters long, and the longest palindromes is 18 letters long at location 14719, 75812, and 173893. Since palindromes less than 10 letters long are not meaningful, they are not included in the data set. The location of the DNA sequence is defined by the order of the DNA of CMV which is 229,354 letters long. For example, the location 200 means the 200th DNA letter in the DNA sequence of CMV.

Data Descriptions

The data set, hcmv.txt, only contains one variable and 296 observations, and it demonstrates the location of palindromes found in the DNA sequence of CMV. To be specific, the mere variable is the location, and 296 observations are the 296 palindromes which are greater than 10 letters long.

Variable Name	Type	Description
Location	Numerical Discrete	The DNA sequence location of Palindromes

Table 1. Dataset variable types and descriptions

Investigations

Since the appearance of palindromes is the same as the distribution of the Poisson Process, in this project, we are going to use the Poisson Process in all analysis. In other words, in this project, we use the Poisson Process as the probability model. Therefore, we emphasize on spacing between each “hit”, appearances of palindromes, and we will divide data into intervals to look at the distribution.

Random Scatter

In order to know whether there are clusters for palindromes, we use the hypothesis test. To be specific, the null hypothesis is that there is no cluster in the data, and the alternative hypothesis is that there are clusters in the data.

Therefore, in the first part, we firstly generate three simulated samples which are uniformly distributed random scatters, and they are in favor of the null hypothesis. Then, we use the observed data to compare with the simulated data, random scatters. Specifically, in order to get a more precise result, we generate three uniformly distributed samples with the same sample size as the data set, 296 palindromes, and in the same DNA sequence, 229,354 base pairs long.

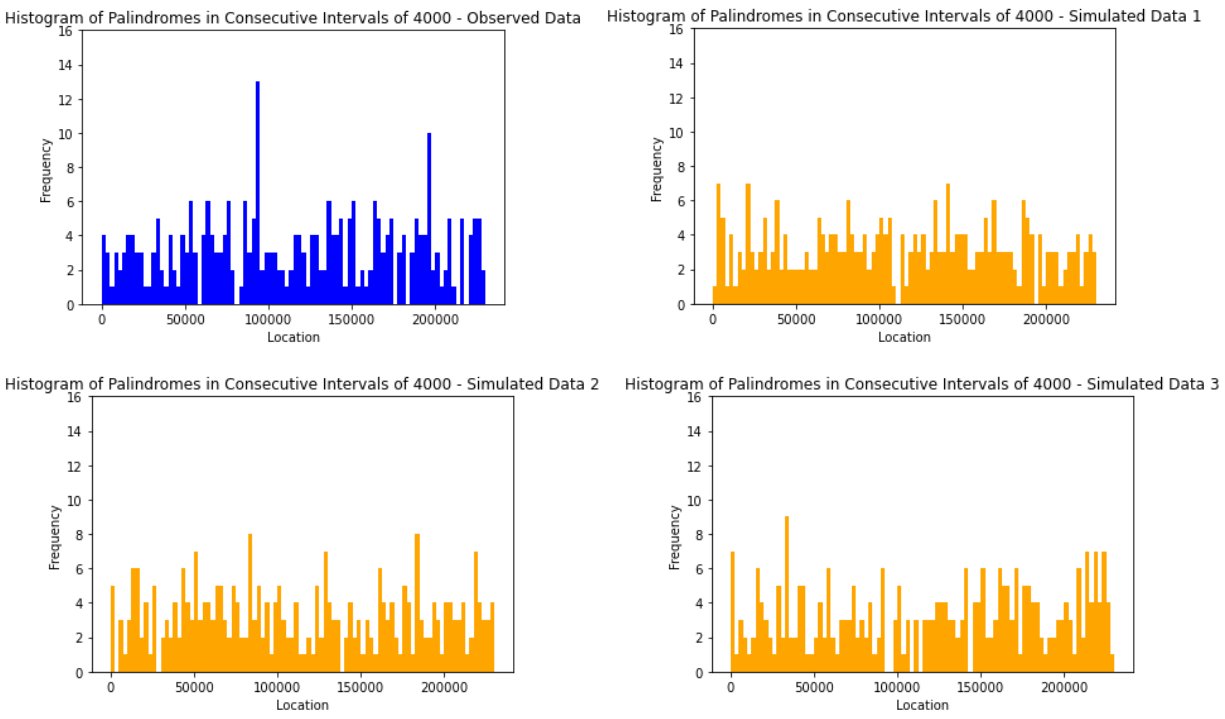


Figure 3. Histograms of locations of palindromes in intervals of length 2500. First histogram is the plot for observed data, and the other three histograms are the plot for simulated data 1 to 3 respectively.

After generating the simulated data, random scatters, we got four histograms of locations of palindromes. The first histogram, blue, is the histogram of observed data from the data set, and the other three histograms, orange, are the histogram made by the simulated data 1 to 3 respectively. Specifically, we divide locations into 93 intervals of length of 2500 base pairs, and the histograms show the frequencies of palindromes in each interval. Meanwhile, the y-axis represents the frequency of the palindromes within the interval, and the x-axis shows the location of the intervals.

Furthermore, if there is an obvious difference between the first histogram and the other three histograms, the test will support the alternative hypothesis. Otherwise, if there is no distinct difference between the histogram of observed data and the ideal histogram, the test will support the null hypothesis. Furthermore, for two histograms and their data, there are 296 palindromes randomly scattered along 229,354 base pairs. The bins of two histograms are both 2500 base pairs long, and, in other words, locations are grouped into intervals of 2500 base pairs.

Upon examination of two histograms, it is really difficult to find an obvious pattern and different between two histograms. Nevertheless, around location 90,000 and 190,000 in the histogram of observed data, there are two distinct intervals of locations containing a relatively high number of palindromes to compare with the rest of the data in all histograms. These are potentially the unusual clusters, but we need more testing for that.

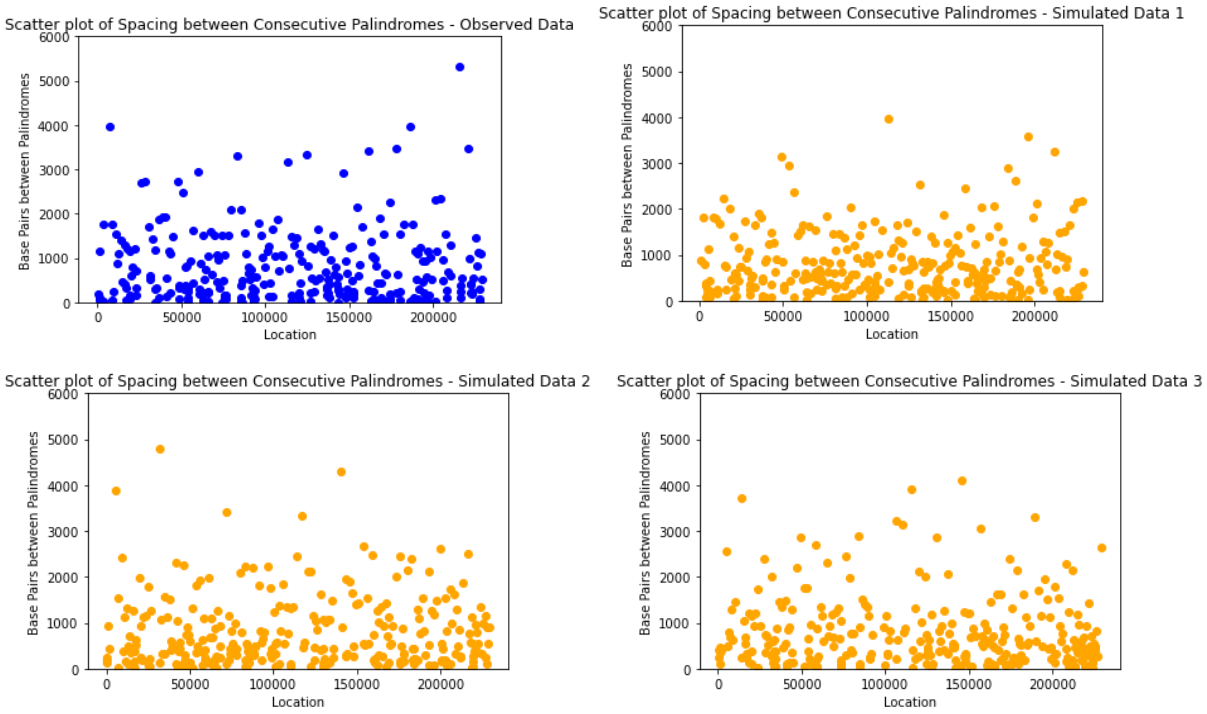


Figure 4. Scatter plots of spacings between palindromes. First plot is the plot for observed data, and the other three are the plots for simulated data 1 to 3 respectively.

We made the scatter plots for spacings between consecutive palindromes above. To be specific, the first plot, blue, is the plot of data from the data set, and the other three plots, orange, are the plots for simulated data 1 to three respectively. Specifically, the spacing is calculated by two consecutive palindromes. The y-axis represents the spacing between the according palindrome and the previous one, and the x-axis indicates the location of the palindromes.

As we can see, there is also no obvious difference between the first plot and other three plots. Specifically, since there is no obvious pattern or clusters, we cannot generally find whether there are clusters or not by the spacings. Then, we will take a look at the counts of palindromes in non-overlapping intervals of length of 2500.

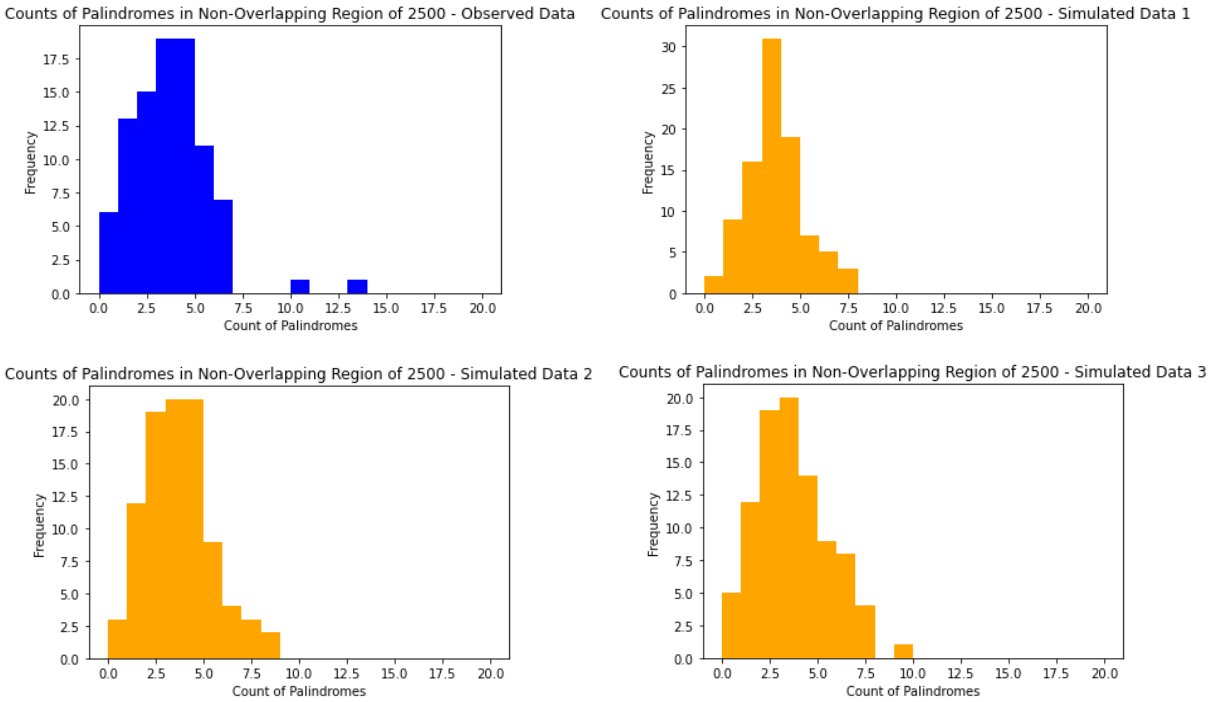


Figure 5. Histogram of Counts of Palindromes in Non-Overlapping Regions of Length 2500. First plot is the histogram for observed data, and the other three are the histograms for simulated data 1 to 3 respectively.

We created histograms for counts of palindromes in non-overlapping regions of length of 2500 base pairs. The blue one is the histogram for observed data, and the other three are histograms for the random scatters. Furthermore, the y-axis represents the frequency of the certain count of palindromes within the interval, and the x-axis represents the counts of palindromes within intervals.

By looking at the counts of palindromes in non-overlapping intervals, we can identify two outliers which have 10 and 13 palindromes in intervals [92,500, 95,000) and [19,500, 197,500) respectively. Specifically, to compare the observed data with the random scatter by the counts of palindromes, there are two obvious outliers, 10 and 13 palindromes in intervals [92,500, 95,000) and [195,000, 197,500) respectively. To compare with the potential unusual clusters by comparing locations of palindromes, they are the same intervals. Therefore, these two intervals are potentially the unusual clusters.

In conclusion, by comparing the observed data with random scatters through locations of palindromes, spacings between palindromes, and counts of palindromes, we can conclude that the data seems to depart from the random scatter. To be specific, although the comparison of spacings between palindromes seems no obvious difference, the other two comparisons show some difference between the data and the random scatters. Therefore, we can say that the data appears to depart from the random scatters. In other words, it implies the existence of unusual clusters of palindromes, and we can identify the origin of replication by finding the unusual clusters. However, we cannot get the real results from the

simple comparisons, and we need more statistical tests to determine whether there are unusual clusters. Thus, in the following investigations, we are going to analyze more detailedly for the question.

Locations and Spacing of Palindromes

For this part, we split the DNA sequences into 50, 55, and 60 intervals respectively. We will investigate whether the distribution of palindromes follows the uniform distribution. In order to get a precise result, we will use graphical methods and goodness of fit tests.

Location of Palindromes

Graphical Method

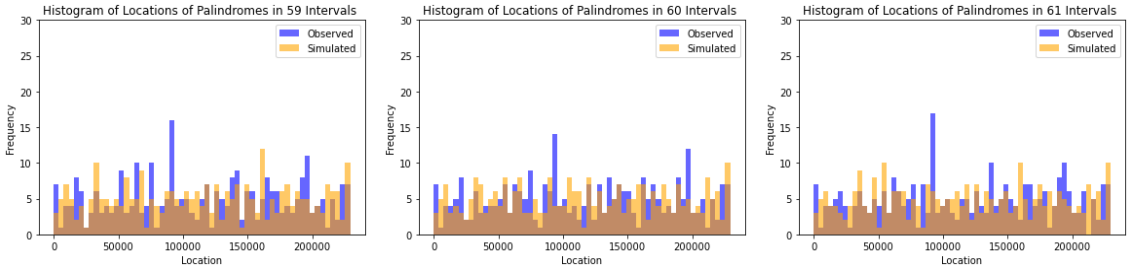


Figure 6. Histogram of locations of palindromes in different numbers of intervals.

We use overlaid histograms as our graphical method to determine whether there are unusual clusters. Specifically, we divided data into different numbers of intervals, 59, 60, and 61 for sensitivity tests. Since we do not want our results to be easily changed by the number of intervals, we need to make sure the results do not change much if we change the number of intervals slightly. On the histograms, the orange bars represent the one of the random scatter generated, and the blue bars are the data observed in the actual DNA sequence of CMV.

Up examination of the histograms above, we can observe that there are two obvious clusters in observed data to compare with the theoretical distribution. To be specific, around location 90,000 and 190,000, there are two intervals which have a relatively high number of palindromes. In other words, these two intervals might be unusual clusters. For sensitivity tests, since we generate histograms for 15, 20, and 25, we can get similar results from all three histograms. In other words, the results do not change much if we change the length of intervals. All in all, there seems to be two unusual clusters around location 90,000 and 190,000 through graphical method.

Chi-square Goodness of Fit Test

For the goodness of fit test for locations, we will test for whether the distribution location of the palindromes is uniformly distributed. Accordingly, the null hypothesis is that the locations of palindromes have uniform distribution, and, thus, the alternative hypothesis is that the locations of palindromes do not have uniform distribution.

For 59 intervals:

T = 91.885, p-value = 0.003

For 60 intervals:

T = 79.000, p-value = 0.042

For 61 intervals

T = 85.662, p-value = 0.016

Upon examining the results of the Chi-Square Goodness of Fit tests, we can conclude that we can reject the null hypothesis with a level of test 5%. Specifically, if we want to reject the null hypothesis, we need our p-value to less than 0.05. As we can see above, the p-values of three different numbers of intervals are all less than 0.05. In other words, we can reject the null hypothesis, and, then, we can conclude that the location of palindromes does not have uniform distribution.

Spacings between Consecutive Palindromes

For spacings between consecutive palindromes, we are going to test whether the spacing between two consecutive palindromes has exponential distribution. Therefore, our null hypothesis is the spacing between two consecutive palindromes has exponential distribution, and, accordingly, the alternative hypothesis is that the spacing between consecutive palindromes.

Firstly, we calculated the λ for exponential distribution, it is $\lambda = \frac{1}{\mu}$. In this question, λ is the inverse of the mean of spacings between consecutive palindromes. After calculating by computer, using Python, we got $\lambda = 0.00129$.

Graphical Method

Then, we can simulate a pseudo set of spacings between consecutive palindromes which has exponential distribution with $\lambda = 0.00129$. Then, we compare the distribution between the observed and simulated data.

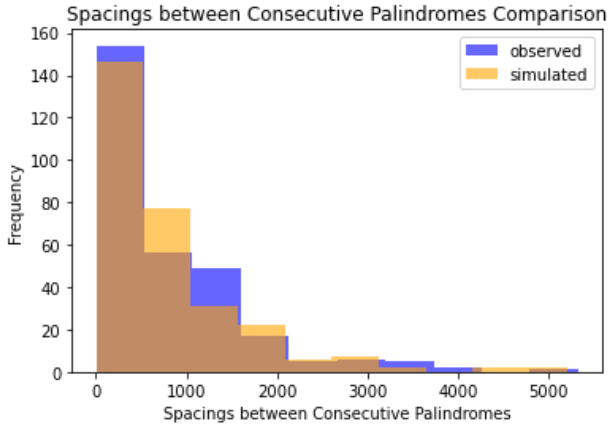


Figure 7. Histogram of spacings between consecutive palindromes compare with the simulated exponential distributed sample.

According to the histogram above. We cannot really identify whether there is a great difference between the pseudo data and the observed data, since they are almost the same although there are some differences for spacings between consecutive palindromes of 1400, and the maximum of the observed data is relatively greater. Thus, we believe that we need a statistical test for that, and we performed the Chi-Square Goodness of fit below.

Chi-Square Goodness of Fit

For this part, we test for if the spacings between consecutive palindromes have exponential distribution. In order to have numbers in all bins, we let the number of intervals to be 7.

T = 85.078, p-value = 3.18e-16

As a result, we can conclude that the null hypothesis can be rejected, and the spacing between consecutive palindromes does not have exponential distribution. To be specific, since the p-value is significantly less than the significance level 0.05, we can reject the null hypothesis. In other words, the spacing between consecutive palindromes does not have exponential distribution.

Spacings between Consecutive Pair of Palindromes and Triplet of Palindromes

In this part, we are investigating the spacings between pairs of palindromes. To be specific, since pair of palindromes and triplet of palindromes both have Gamma distribution as $Gamma(2, \lambda)$ and $Gamma(3, \lambda)$ respectively. Thus, we are testing for whether spacings between consecutive pair of palindromes has $Gamma(2, \lambda)$ distribution whether spacings between consecutive triplet of palindromes has $Gamma(3, \lambda)$ distribution.

Similarly, we calculated the λ for pair and triplet parts respectively, and we got 0.000645 for pair part and 0.000429 for triplet part.

Graphical Method

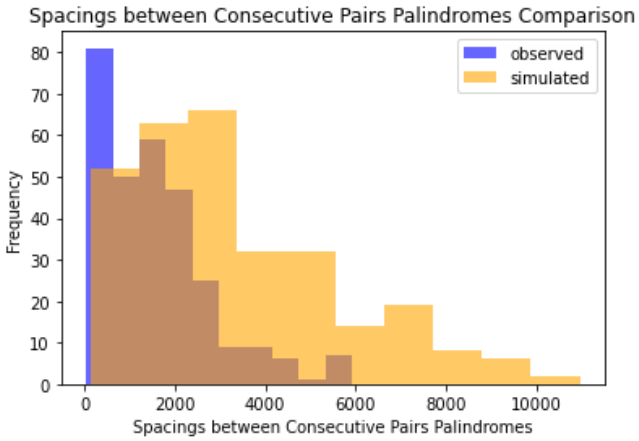


Figure 8. Histogram of spacings between consecutive pairs of palindromes compare with the simulated Gamma distributed sample.

According to the histogram above, we can easily identify the difference between the observed data and the simulated data. In other words, the observed data does not have the Gamma distribution. In other words, the spacing between pairs of palindromes does not have Gamma distribution.

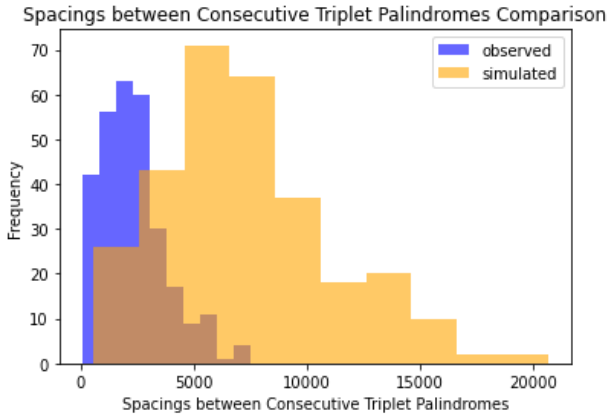


Figure 9. Histogram of spacings between consecutive triplet of palindromes compare with the simulated Gamma distributed sample.

Similarly, according to the histogram above, we can easily identify the difference between the observed data and the simulated data. In other words, the observed data does not have the Gamma distribution. Therefore, the spacing between triplets of palindromes does not have Gamma distribution. However, we need a hypothesis test to precisely determine the results, and we performed the Chi-Square Goodness of Fit test below.

Chi-Square Goodness of Fit

For pairs and triplets, we use 16 and 19 intervals to use chi-square goodness of fit tests since we want to keep there is at least one in each bin.

For Pairs:

T = 2448.132, p-value = 0.000

For Triplets:

T = 2991.366, p-value = 0.000

Upon examining the results we got, the p-values for both paris and triplets are extremely small. In other words, since p-values are both less than the significance level 0.05, we can reject the null hypothesis. In other words, the distributions of spacing between pairs and triplets of palindromes are both not in Gamma distribution.

Conclusion

By investigating the locations, spacings between consecutive palindromes, and spacing between pairs and triplets of palindromes, we can conclude that the distribution of observed palindromes is not distributed as expected. Therefore, we can conclude that there are unusual clusters in the CMV's DNA sequence.

Counting Palindromes in Various Regions of the DNA and Sensitivity Analysis

Graphical Summaries

Recall that the CMV DNA in our data is 229,354 letters long, and in our dataset, we have 296 observations, which denote the locations of the palindromes which were discovered to be at least 10 letters long. The next method we provide involves dividing the DNA string into equal-length intervals. We will also be comparing our bin distributions with a fitted Poisson distribution. The amount of intervals we split the dataset into are as follows: 20, 25, 50, 75, 100, 150. After binning our DNA strand, we create a count dataset for each number of intervals. In particular, we end up with a frequency distribution of the palindromes in each interval. Using the mean counts, we generate a best-fitting Poisson model for each frequency distribution. Fitting the Poisson distribution to count data is necessary as it is discrete and positive, and it aligns with the underlying assumption of a random process of drawing a small number of events in intervals of time.

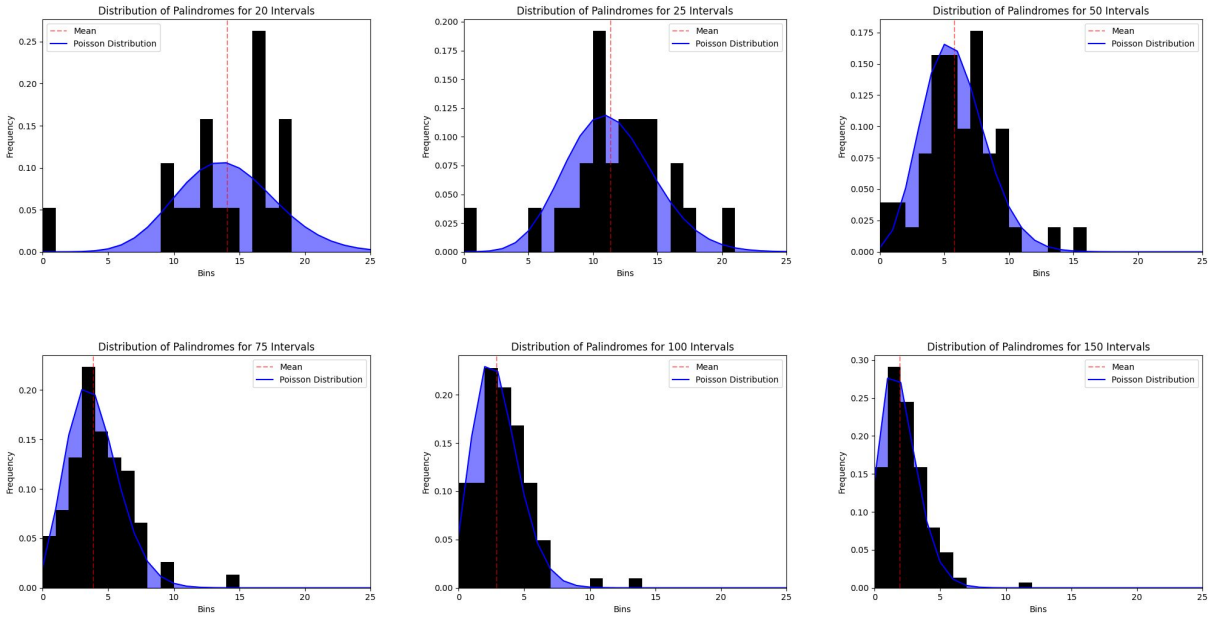


Figure 10. Distributions of palindromes fit to Poisson distribution for DNA splits into 20, 25, 25, 75, 100, and 150 intervals, respectively.

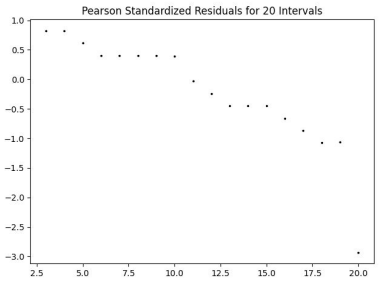
As we split the DNA string into more intervals, the observed distribution begins to resemble the Poisson distribution. We visualize this convergence by plotting the Pearson residuals [1], and can infer a Poisson structure in the data if the absolute values of the lengths of the residuals are less than or equal to three. We use Pearson residuals to correct for unequal variance in the raw residuals by dividing by the standard deviation. We define the Pearson residual procedure in the Theory section.

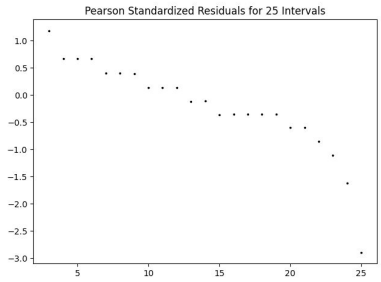
Before presenting the Pearson residuals, we note the differences in skewness between the observed distributions and the estimated Poisson distributions for each number of intervals, caused by the occurrences of outliers mostly after bins 10 or 15, solely from the figure above.

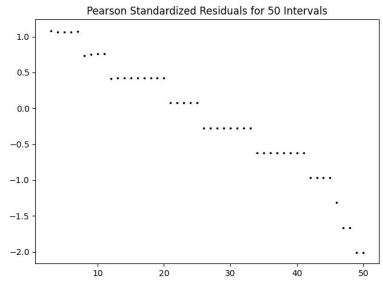
Number of Intervals	Observed	Poisson PMF
20	-0.9	0.79
25	-0.52	1.17
50	0.51	2.57
75	1.12	3.67
100	1.47	4.64
150	1.57	6.36

Table 2. Skew values for the observed data and PMF with 20, 25, 50, 75, 100, and 150 equal-length intervals.

We infer from the table above that the fitted Poisson PMF is positively skewed, and both distributions become more so as we create more intervals. However, given that the positive skews of the fitted PMFs are greater in magnitude than those of the observed distributions, the PMF underestimates the expected number of palindromes at a given location due to its smaller weight on outliers. The table and graphical methods tell us that the distribution of palindromes at different locations follows a Poisson distribution and is slightly positively skewed with outliers. Below, we display the Pearson Standardized Residual plots for each number of intervals.







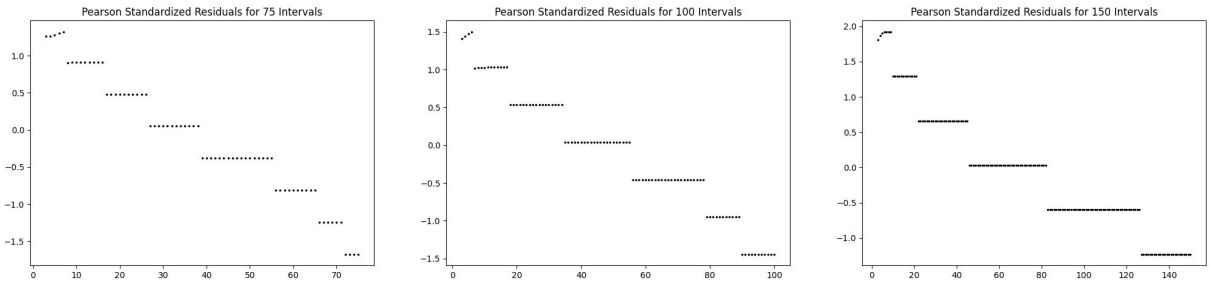


Figure 11. Pearson standardized residuals per number of intervals.

Since the magnitude of the standardized residuals never exceeds 2, we can say that it is very likely the binned palindrome counts follow a Poisson distribution.

Cluster Analysis

In this section, we use randomization to examine the largest cluster of palindromes in a sub-interval. Our threshold is $\alpha = 0.05$. Under the Poisson process model, we know that the quantity of palindromes in an interval are independent observations, from a Poisson Distribution. Similarly, we see the interval that contains the most palindromes will be the maximum of these independent observations. We are able to approximate the probability of k being our largest observation (cluster size), using our (lambda-hat) value. In the table below, shows 4 different interval amounts: (2500,3500,5000,7000). We observe that as the cluster sizes vary as a result of the different interval lengths. With our threshold α we also observe that our p-value, which is the largest cluster size expected from the Poisson Process, changes as well. Furthermore, from our data, we see that the p-values for 2500, 3500,are statistically significant, as their corresponding p-values show that there is evidence that the largest clusters within these intervals are larger than what we would expect using the Poisson Process. While the interval of 5000 appears to be statistically significant as well, the interval lengths of 2500 and 3500 appear to have more significance in this instance.

Interval Length	P-Value	Lambda Value (lambda-hat)	Max Cluster Size in Interval	Subintervals
2500	8.490572e-86	3.22	13	92
3500	1.397445e-216	3.48	16	66
5000	8.223000e-44	6.43	18	46
7000	1.460012e-01	8.97	18	33

We also generated a random sample with uniform distribution, then divide the sample and our own data into different numbers of intervals, which are 30, 50, 70, 100, 150, 200, 500, 2000, 3000. Then get the probability of biggest count in the intervals of the random sample is greater than the biggest count in the intervals of our data, the probabilities shows below:

# of Intervals	30	50	70	100	150	200	500	2000	3000
probability	0.9995	0.686	0.5415	0.847	0.567	0.06	0.1405	0.004	0

According to the result we got, although we have some exceptions, the overall probability decreases as the number of intervals gets bigger. When we set the number of intervals to be 30, indicating that a single interval's size is relatively large, we got a probability close to 1, which implies that we can hardly distinguish any clustering. On the contrary, when we set the number of intervals to over 500, the probabilities become too small which means that there is too much clustering and would decrease the accuracy of finding the origin replication. Therefore, we conclude that the best number of intervals for looking for the pattern of clusters would be about from 70 to 200, since it can optimize the chance of finding the overlaps in these observations.

Theory

Goal

Our **goal** of this project is to find the origin of replication of the CMV, and identifying the unusual clusters is a way to find the origin of replication. Therefore, in this project, we will find the unusual clusters by hypothesis testing and the Homogeneous Poisson Process. Specifically, since the data seems suitable for Homogeneous Poisson Process, and the Homogeneous Poisson Process is the model for random scatter, uniformly distributed. Thus, we can use hypothesis testing to test whether there is significant difference between the data and the simulated random scatters, and we can identify the unusual clusters to find the origin of replication.

Homogeneous Poisson Process

We use the **Homogeneous Poisson Process** to analyze the distribution of palindromes in this project. To be specific, the homogeneous Poisson Process is a natural model for points randomly distributed on a line, such as arrival times. Furthermore, the palindrome in the DNA chain is like the points on the line. Specifically, the points are the appearances of palindromes, and the line is the DNA sequence chain. In other words, the Poisson Process describes the distribution of palindromes in the DNA sequence. There are three major features of the Homogeneous Poisson Process:

1. The underlying rate, λ, at which points/hits/occurs, does not change with the line/location/time
2. The number of points falling in separate regions are independent
3. No two points can land in exactly the same place

Meanwhile, the distribution of palindromes in the DNA sequence are also follow these three features, because:

1. The poison process is a good reference model for making comparisons because it is a natural model for uniform random scatter
2. The DNA sequence can be thought of as a line, and the location of a palindromes can be thought of as a point on the line
3. Palindromes are scattered randomly and uniformly across the DNA
4. The number of palindromes in any small piece of DNA is independent of the number of palindromes in another, non-overlapping space
5. The chance that one tiny piece of DNA has palindrome in it is the same for all tiny pieces of the DNA

Moreover, the probability mass function of Poisson distribution is $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, but typically λ is unknown. Thus, the estimator for λ is the observed average number of counts per interval. Then, we define the Homogeneous Poisson Process as $\{N(t),\ t > 0\}$ where t is a non-negative integer random variable. Then, the λ for the Poisson Process is given by λt. Therefore, we have the probability model of the Poisson Process:

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

In this project, t is the location in the DNA sequence, N(t) is the number of palindromes and t is the locations of the DNA sequence. For example, the first palindrome appears at the location 177, so $N(177) = 1$.

Pearson Residuals

Before we define the **Pearson residual**, let us recall the idea of the **raw residual**. Raw residuals represent the difference between the estimated value of a Poisson model and the observed value. Poisson random variables assume that the mean is equal to the variance, so the variances of the residuals are unequal. Analytically, we define the raw residual as:

$$r_i = y_i - e^{\{X_i \beta)}$$

In order to correct the unequal variance, we divide r_i by the standard deviation, and derive the following equation for the Pearson residuals:

$$p_i = \frac{r_i}{\sqrt{\widehat{\varphi} e^{\{X_i \beta\}}}}, \text{ where } \widehat{\varphi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - e^{\{X_i \widehat{\beta}\}})^2}{e^{\{X_i \widehat{\beta}\}}} \text{ is a dispersion parameter}$$

Hypothesis Testing

In this project, we will use hypothesis tests to test whether there are unusual clusters and whether the Poisson Process fits our data. Specifically, we test the presence/absence of unusual clusters with multiple hypothesis tests, and we will use goodness of fit to test whether the Poisson Process fits our data. For testing the presence/absence of unusual clusters, our null hypothesis is that the unusual clusters do not exist, and, accordingly, our alternative hypothesis is that the unusual clusters exist. Then, for the goodness of fit, our null hypothesis is there is no significant difference between the data and the Poisson Process,

and, thus, our alternative hypothesis is that there is significant difference between the data and the Poisson Process.

Chi-Squared Goodness-of-Fit Test

The Chi-Squared Goodness of Fit test is a Parametric hypothesis test, which allows us to determine whether or not a certain variable is likely to come from a specific distribution. We use this test in order to check whether our Poisson Process Model suits our data or not. Additionally, we are able to use the test to determine whether or not our sample is a good representation of the population.

For this test, we must begin with a null hypothesis stating that our data follows *some specific distribution*. In this case, if we were to reject the null hypothesis, this would suggest that there is an abnormality in the distribution of palindromes.

P-Value

The p-value is the probability of obtaining a result that is at least as extreme as the observed results under the null hypothesis. As such, a very small p-value means the extreme result would be unlikely to occur under the null hypothesis. Typically, a p-value of less than 0.05 is considered extreme enough to reject the null hypothesis as being too unlikely.

Given an observed test statistic t from an unknown distribution T,

$$\text{P-value} = P(T \geq t) \text{ for a right tailed test}$$
$$\text{P-value} = P(T \leq t) \text{ for a left tailed test}$$
$$\text{P-value} = P(|T| \geq |t|) \text{ for a two sided test}$$

Exponential Distribution

Exponential Distribution is the probability distribution of the time between events in a Poisson point process model and the continuous analogue of the geometric distribution. It is useful for testing product reliability and modelling waiting times. It takes in λ as the rate parameter of the distribution which supports the time interval of [0, ∞) and x as a continuous random variable.[8]

PDF:

$$f(x;\lambda) = \begin{cases} \lambda e^{-(\lambda x)} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

CDF:

$$F(x;\lambda) = \begin{cases} 1 - e^{-(\lambda x)} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Since we need to investigate whether the palindromes in the CMV's DNA sequence are distributed as expected, we need to use exponential distribution. Since for the hypothesis that there are no unusual clusters in the DNA locations, the spacing between consecutive palindromes are distributed as Exponential distribution. Therefore, if we can determine that the spacing between consecutive

palindromes have no exponential distributions, we can conclude that there are some unusual clusters in the DNA sequence.

Gamma Distribution

Gamma distribution, in statistics and probability theory, is a continuous distribution. It is characterized by its' 2 parameters: Shape Parameter k (with constraint $k > 0$), Scale Parameter Θ (with constraint $\Theta > 0$). The Gamma distribution is typically used in order to predict the waiting time of when a k-th event will occur.

The formula is as follows:

$$f(x) = \frac{1}{\Gamma(k)\Theta^{k}} x^{k-1} e^{-x/\Theta} \text{ for } x > 0$$

Additionally, our CDF is :

$$F(x) = \int_{0}^{\infty} \frac{(k, x/\Theta)}{\Gamma(k)} = \frac{(k, \lambda x)}{\Gamma(k)}$$

Since we need to investigate whether the palindromes in the CMV's DNA sequence are distributed as expected, we need to use Gamma distribution. Since for the hypothesis that there are no unusual clusters in the DNA locations, the spacing between pairs and triplets of palindromes are distributed as Gamma distribution. Therefore, if we can determine that the spacing between pairs and triplets of palindromes have no Gamma distributions, we can conclude that there are some unusual clusters in the DNA sequence.

Contributions

Jared Dishman: Theory (P-value, Exponential Distribution)
Alison Camille Dunning: Introduction, Investigation (Small Question 3), Theory (Pearson Residuals)
Ruotian Gao: Investigation (Small Question 4)
Yaoxin Li: Data, Investigation (Small Question 1, Small Question 2), Theory(Goal, Homogeneous Poisson Process, Hypothesis Tests, Exponential Distribution, Gamma Distribution)
Bryan Talavera: Theory (Chi-Squared Goodness-of-Fit, Gamma Distribution), Investigation (Small Question 4)
Abdiaziz Weheliye: Introduction, Chi-Squared Goodness-of-Fit (Small question 3)
Ye Yint Win: Background, Theory (Exponential Distribution)

References

1. 12.3 - Poisson regression. (n.d.). Retrieved February 17, 2021, from https://online.stat.psu.edu/stat462/node/209/
2. En.wikipedia.org. n.d. *Avery–MacLeod–McCarty Experiment* https://en.wikipedia.org/wiki/Avery%E2%80%93MacLeod%E2%80%93McCarty_experiment
3. Crick, Watson, and Franklin: The Race to Discover the Structure of DNA https://www.khanacademy.org/humanities/big-history-project/life/knowing-life/a/crick-watson-and-franklin2

4. Centers for Disease Control and Prevention n.d.*Cytomegalovirus (CMV) and Congenital CMV Infection.* https://www.cdc.gov/cmv/overview.html
5. Mayo Clinic n.d. *Cytomegalovirus (CMV) Infection* https://www.mayoclinic.org/diseases-conditions/cmv/symptoms-causes/syc-20355358
6. Leung, Ming-Ying (April 21, 2005) et al. *Nonrandom Clusters of Palindromes in Herpesvirus Genomes*.Journal of Computational Biology, v12 n3 pg 331-354
7. En.wikipedia.org. n.d. *Gamma distribution*. [online] Available at: <https://en.wikipedia.org/wiki/Gamma_distribution>.
8. En.wikipedia.org. n.d. *Exponential Distribution* https://en.wikipedia.org/wiki/Exponential_distribution