CASE STUDY 3: SEARCH FOR THE UNUSUAL CLUSTER IN THE PALINDROMES

May 11, 2018

Chengyu Chen (A14051607), 2nd Year Applied Mathematics; Data Science, MATH 189
Chenyue Fang (A13686794), 2nd Year Probability and Statistics, MATH 189
Daniel Lee (A13726312), 2nd Year Probability and Statistics; Data Science, MATH 189
Xinran Wang (A13564644), 2nd Year Probability and Statistics, MATH 189
Yuqi Wang (A13532155), 2nd Year Applied Mathematics, MATH 189
Ning Xu (A92061610), 3rd Year Probability and Statistics; Economics, MATH 189

Introduction

The human cytomegalovirus virus (CMV) is a potentially life-threatening disease for people with suppressed or deficient immune systems. For example, this may affect people with organ transplants or HIV. To develop strategies to combat this virus, researchers study the way in which the virus replicates. In particular, they search for a specific location on the virus' DNA that contains instructions for its reproduction: origin of replication. Using this information, they can then design new drugs and treatments by localizing the replication process.

A virus' DNA, thought of as a long, coded message from a four-letter alphabet (A, C, G, T), contains all of the information necessary for it to grow, survive, and replicate. Since the alphabet is so small, DNA sequences contain many patterns, some of which may flag sites of interest on the DNA, such as the origin of replication. A complimentary palindrome is one such pattern. In DNA, the letter A is complementary to T and G is complementary to C, and a complementary palindrome is a sequence of letters that reads in reverse as the complement of the forward sequence. For example, *GGGCATGCCC* is a complementary palindrome.

The origin of replication for two viruses (Herpes simplex and Epstein-Barr) from the same family as CMV, the herpes family, are marked by complimentary palindromes. The Herpes simplex is marked by a long palindrome of 144 letters while the Epstein-Barr virus has several short palindromes and close repeats clustered at the origin of replication. For the CMV, the longest palindrome is 18 base pairs and altogether contains 296 palindromes between 10 and 18 base pairs long. Researchers conjecture that clusters of palindromes in CMV may lead to a possible origin of replication, similarly to the Herpes simplex and Epstein-Barr virus.

To determine whether a particular DNA sequence is the origin of replication, the DNA is cut into segments and each individual segment is tested to determine whether it can replicate. This process can be very time consuming and expensive without leads on where to begin the search. Therefore, a statistical investigation of the DNA is done to identify unusually dense clusters of palindromes which can help narrow the search and reduce the amount of testing needed.

In this paper, we will search for unusual clusters of complementary palindromes. The overarching research question is: “How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site? Based on our analysis, we will then provide recommendations to biologists who are about to start experimentally searching for the origin of replication.

The Data

The full DNA sequence of CMV was published in 1990 by Chee et al. and is approximately 200,000 base pairs long. By using search algorithms to screen the sequence for various types of patterns, in 1991, Leung et al. found a total of 296 palindromes that were at least 10 letters long. The longest palindromes found are 18 letters long and occur in 4 locations along the sequence. Palindromes shorter than 10 letters are ignored.

The dataset used for our analysis consists of a list of the locations of these 296 palindromes along the CMV DNA sequence.

Background

DNA, abbreviation of deoxyribonucleic acid, is the hereditary material in living organisms, which means it carries unique hereditary information of each individual. DNA was discovered in 1896, but its function of carrying hereditary information remained unknown until 1943. In 1953, Francis Crick and James Watson found that the structure of DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone. The four bases are adenine (A), thymine (T), cytosine (C) and guanine (G), and they form code for storing information in DNA. The two kinds of base pairs are the A-T pair and C-G pair. Human DNA has nearly 3 billion base pairs with 99 percent of them are the same in all individuals. DNA can make copies of itself with each strand of DNA being used for duplicating the new sequence. Furthermore, DNA can also be used for transcription of RNA which carries codes from DNA in nucleus to protein (U.S National Library of Medicine).

A virus is composed of a nucleic acid molecule (either DNA or RNA) wrapped with a protein shell called capsid, and the nucleic acid controls its replication. Although viruses carry genetic information, they are not considered as a life form because they do not have cell structure. Viruses cannot live by themselves, so they must replicate inside the host cells of living organisms. In E.coli replication, a snipping enzyme cuts the DNA strand at one point and lots of free nucleotides appear around the point. Then, as one free nucleotide meets its counterpart, they stick to each other while other nucleotides bouncing away. As more nucleotides are added, a clipping enzyme makes them together (Bradic 15).

Human Cytomegalovirus (HCV) is a virus that can infect people of all ages. Once the person is infected with CMV, the virus stays dormant in the body for life and becomes harmful when the virus enters productive cycle, where the illness can reactivate. Although there is nearly no sign or symptom on most people with infection, CMV causes serious illness on people with poor immune system such like transplant patients and AIDS patients. According to statistics research, 10% to 15% of children are infected with CMV before the age of 5, and over half of adults by age 40 are infected. In order to find a vaccine against CMV, researchers need to locate the origin of replication for the virus (Centers for Disease Control and Prevention).

Investigations

Scenario 1: Random Scatter

We will begin our analysis by investigating the structure of the data in terms of departures from a random scatter of palindromes across the DNA. It is important to note that a uniformly random scatter does not necessarily mean that palindromes will be exactly equally spaced. Nevertheless, most random scatters will lack a distinctive pattern. In the following visualizations, the structure of the original data is compared with the structure of three monte carlo simulated random scatters from the uniform distribution. In particular, 296 palindromes were randomly scattered along a DNA sequence of 229,354 base pairs. Multiple random scatters were generated for the sake of replicability. The locations of palindromes, the spacing between palindromes, and the counts of palindromes in non-overlapping regions of the DNA are compared across the original data and the random scatters.

In the following visualizations, the original data is shown in red while the three random scatters are each shown in black.

First, the locations of palindromes are compared. *Figure 1* depicts a simple dot plot of the palindrome locations. From this visualization alone, it is difficult to recognize any patterns in either the original data or the random scatters. However, we can observe that the palindrome locations are not exactly equally spaced, as discussed earlier.

Figure 1. Simple Dot Plot of Palindrome Locations.

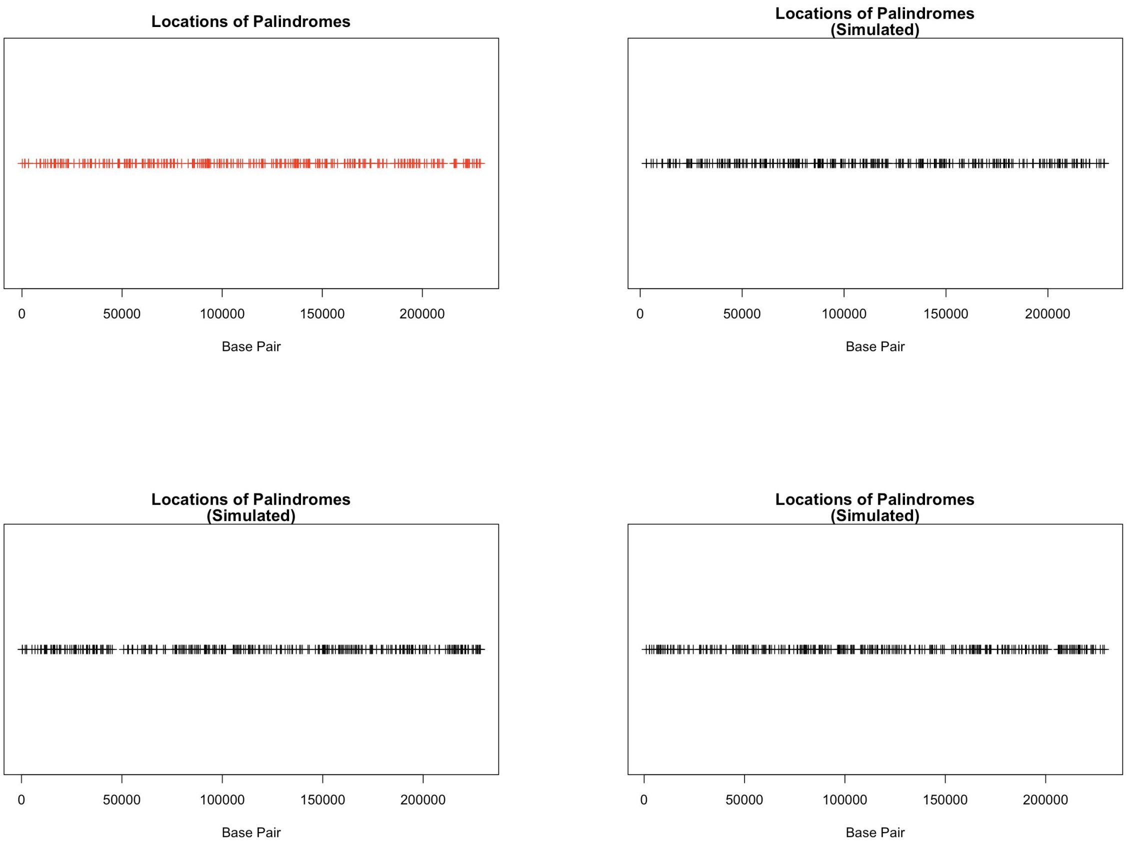
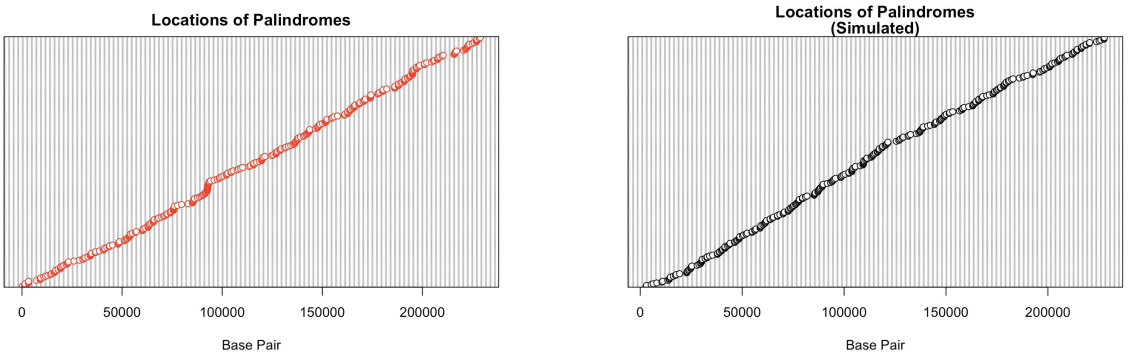


Figure 2 depicts a modified dot plot of the palindrome locations with the x-axis representing base pair location and the y-axis representing the palindrome location order statistic units. From this visualization, palindrome locations from the original data appear to be clustered around the 90000th base pair. Clusters are not as apparent in any of the random scatters, indicating the data's departure from a random scatter.

Figure 2. Modified Dot Plot of Palindrome Locations



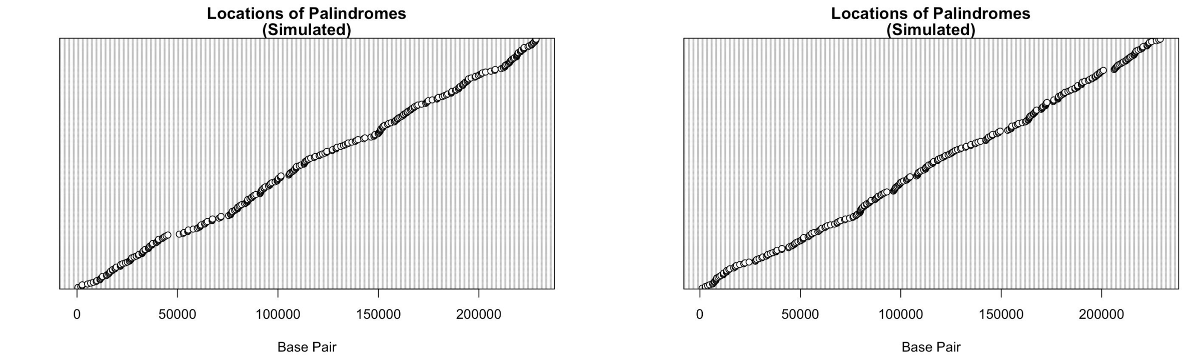
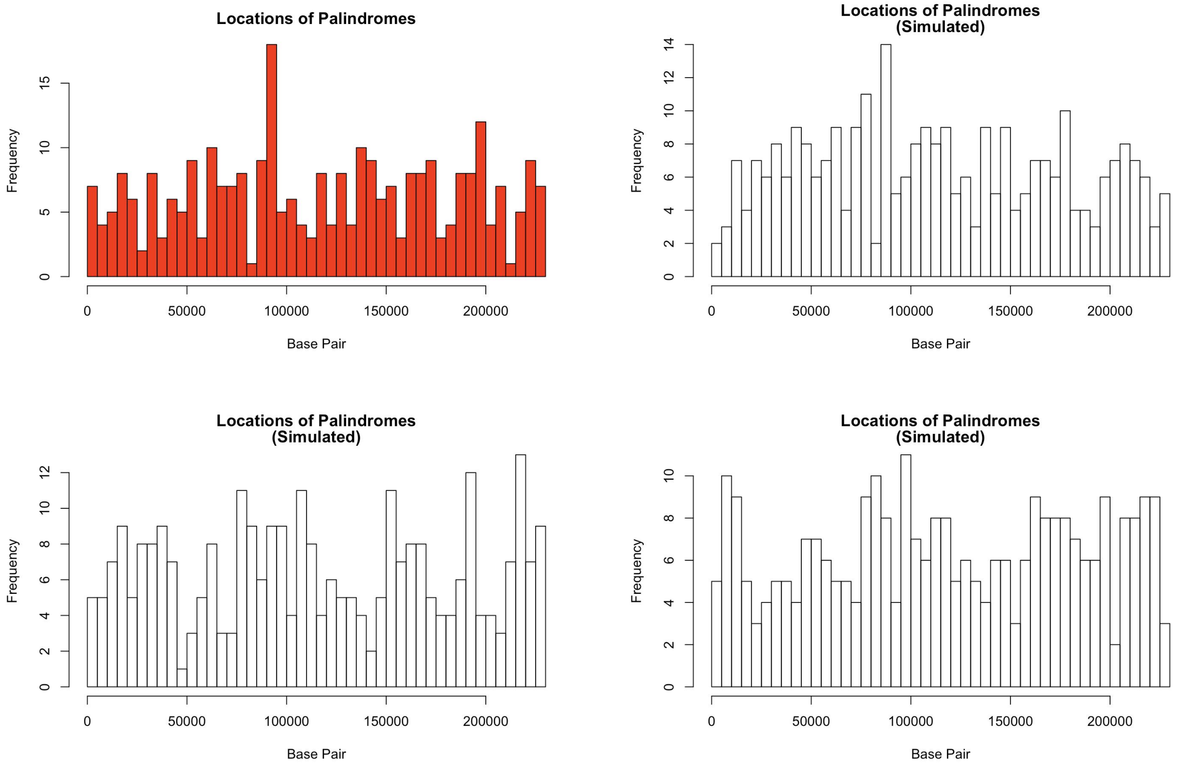


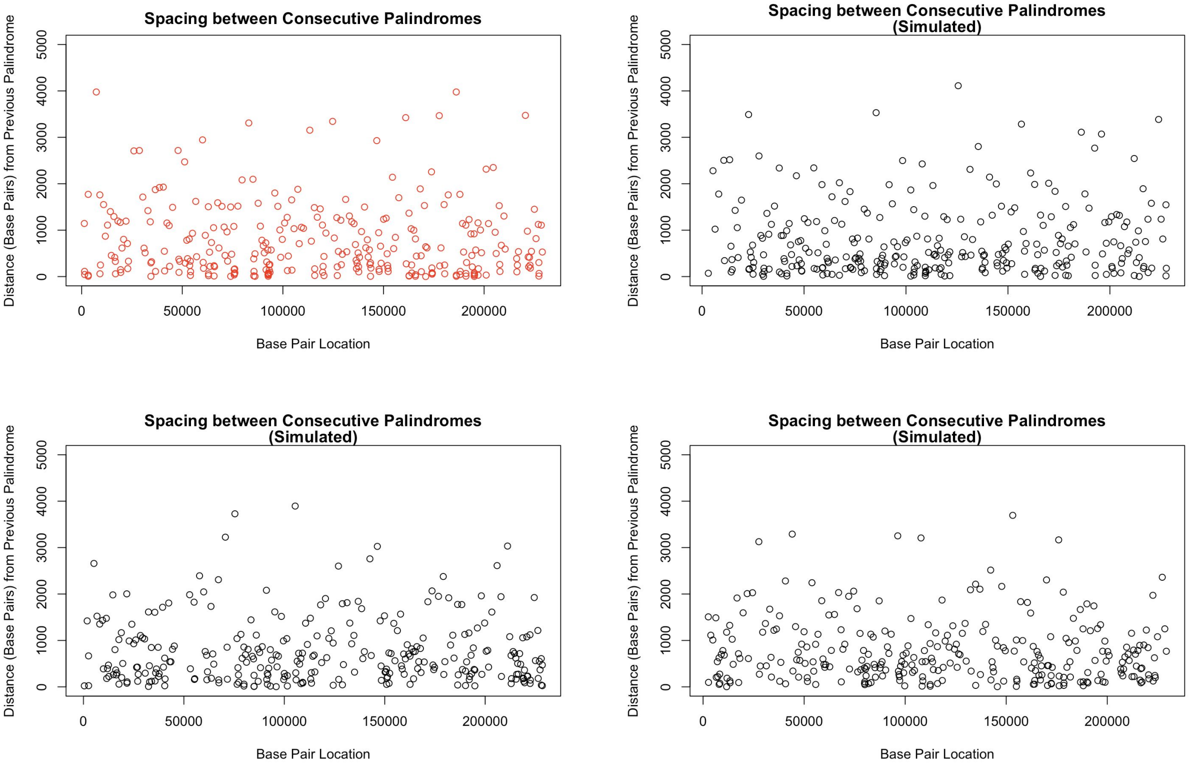
Figure 3 depicts a histogram of the palindrome locations. From this visualization, it is further apparent that palindrome locations from the original data appear to be clustered around the 90000th base bair. Again, clusters are not as apparent in any of the random scatters, indicating the data's departure from a random scatter.

Figure 3. Histogram of Palindrome Locations



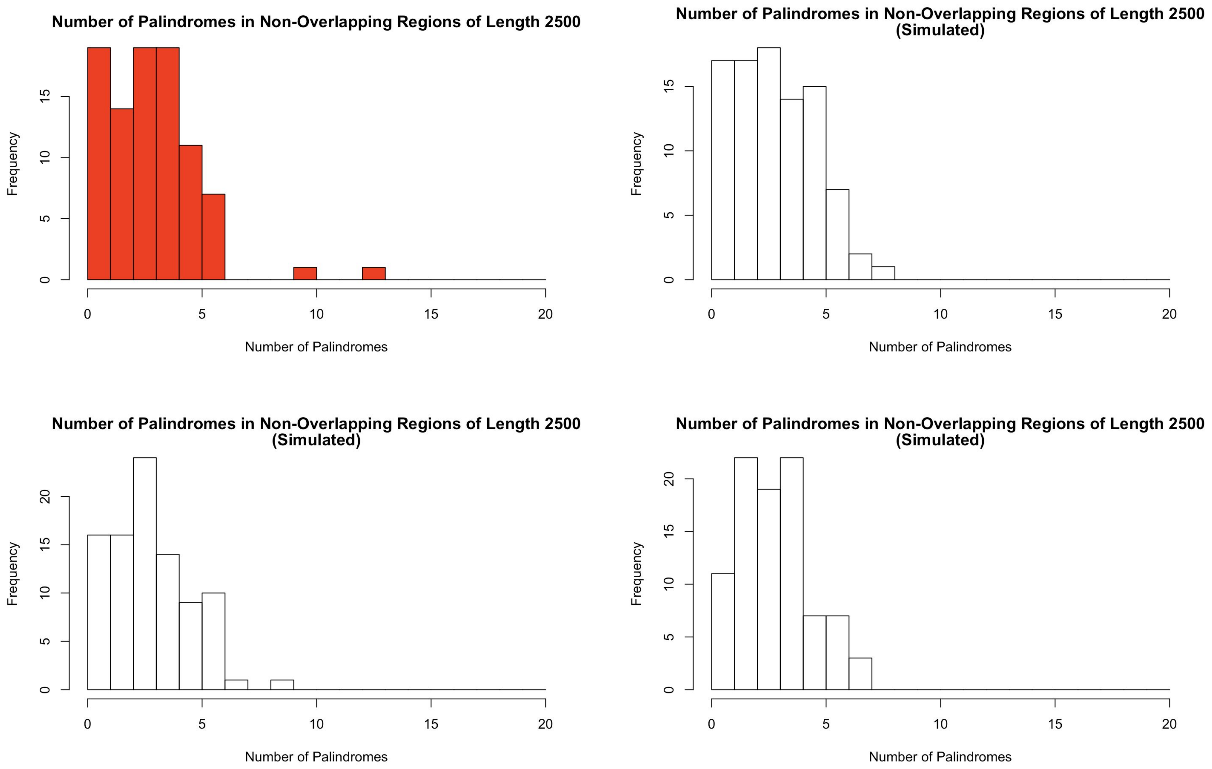
Second, the spacing between consecutive palindromes are compared. *Figure 4* depicts a scatterplot of spacing between consecutive palindromes. Similarly to *Figure 1*, from this visualization alone, it is difficult to recognize any patterns in either the original data or the random scatters. In following scenarios, we will further investigate the spacing between consecutive palindromes to determine if the data departures from a random scatter.

Figure 4. Scatterplot of Spacing Between Consecutive Palindromes



Finally, the counts of palindromes in non-overlapping regions of the DNA are compared. *Figure 5* depicts a histogram of the number of palindromes in non-overlapping regions of length 2,500 base pairs. From this visualization, there appears to be outliers in the distribution's right tail of the original data, possibly corresponding to the cluster discussed earlier around the 90000th base pair. Outliers are not as apparent in any of the random scatters, indicating the data's departure from a random scatter.

Figure 5. Histogram of Number of Palindromes in Non-Overlapping Regions



Therefore, after the locations of palindromes, the spacing between palindromes, and the counts of palindromes in non-overlapping regions of the DNA were compared across the original data and the random scatters, the data appears to departure from a random scatter. This may indicate the existence of clusters of palindromes, hinting at a potential replication site. This possibility will be further investigated in the following scenarios using graphical methods and formal statistical tests.

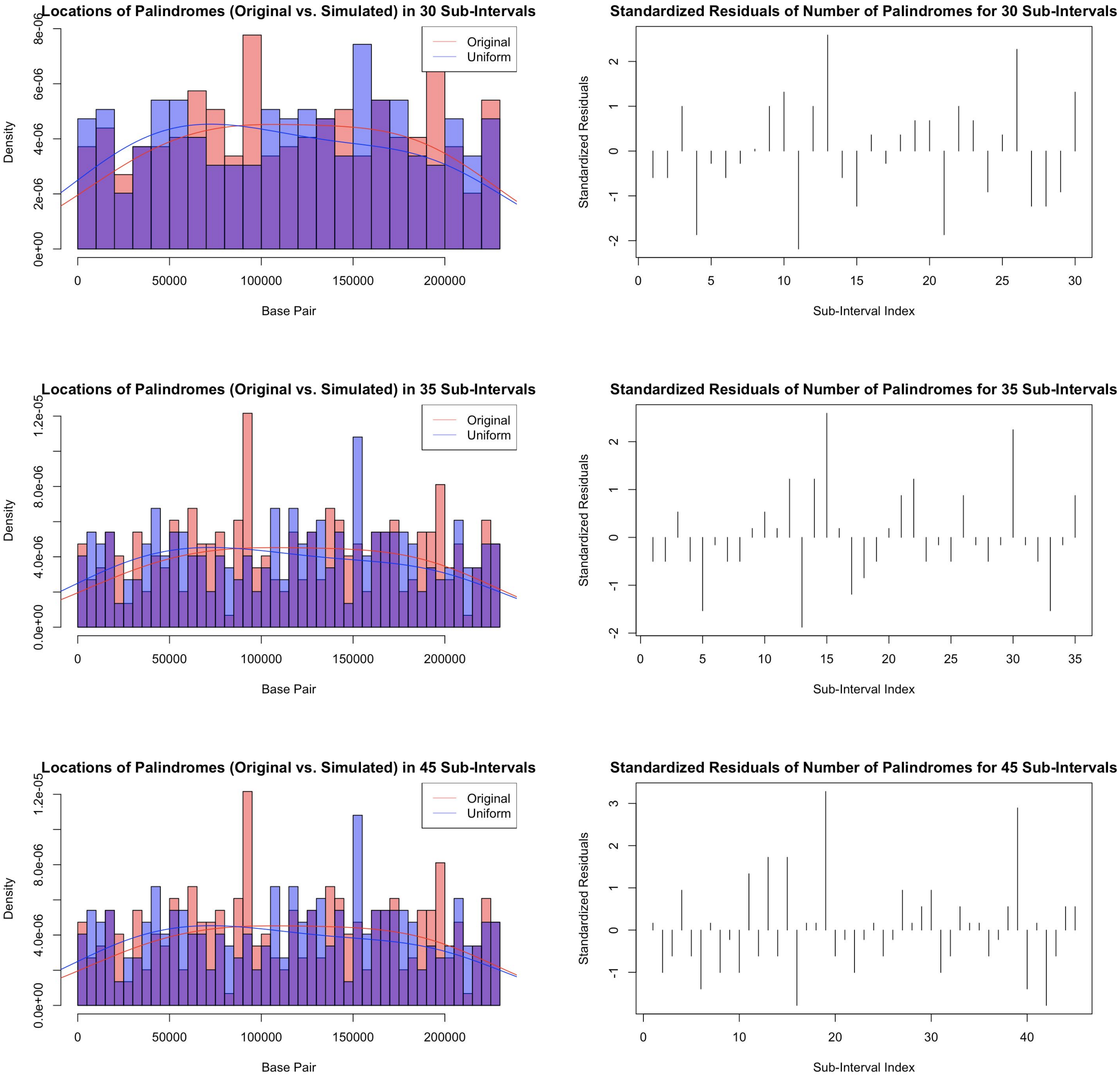
Scenario 2: Locations and Spacings

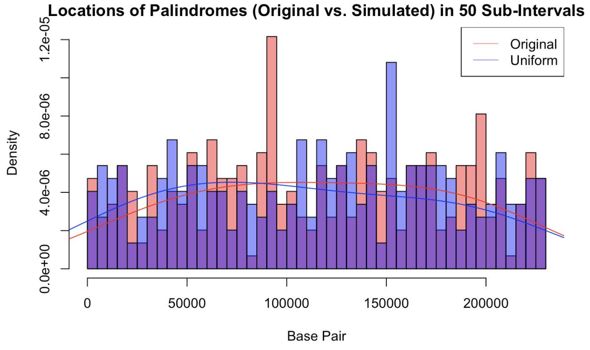
The entire dataset is divided into 30, 35, 45, 50, 55, and 59 sub-intervals for 6 times of testing. We want to investigate how the distribution of the locations of palindromes in each interval fits uniform distribution. In order to achieve our goal, we used graphical methods, goodness of fit testing statistic, and standardized residuals to examine the differences between the two distributions. We specifically selected values of k under 59 since we wanted to ensure that the expected value for each bin is at least 5.

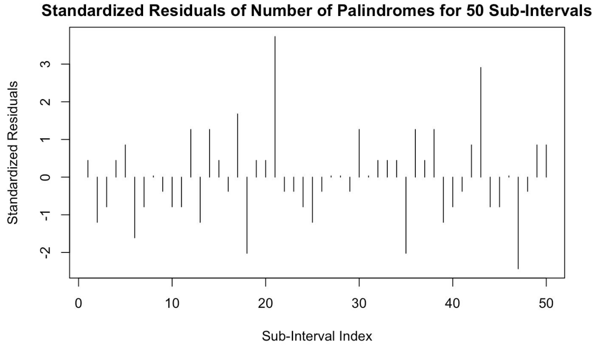
Graphical Methods

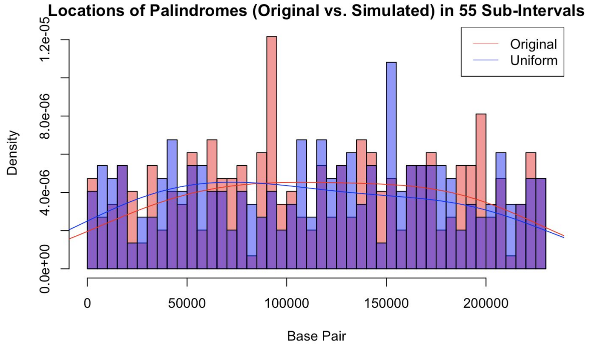
Figure 6 depicts histograms of the palindrome locations in non-overlapping regions of length 7,647 base pairs, 6,555 base pairs, 5,098 base pairs, 4,589 base pairs, 4,141 base pairs, and 3,889 base pairs. In this visualization, the red graph represents the actual data, and purple represents generated uniform distribution. From the histograms, the frequency of counts in each bin varies largely, and it is apparent that the skewness of the density curves are different. Therefore, it is reasonable to question whether the actual data is uniformly distributed, and further if the actual data contains unusual clusters of palindromes.

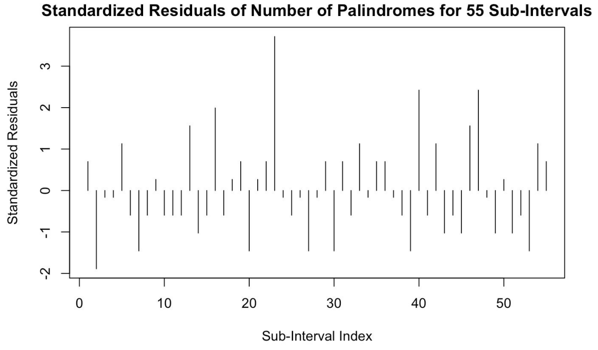
Figure 6. Histogram and Residuals of Number of Palindromes in Non-Overlapping Regions

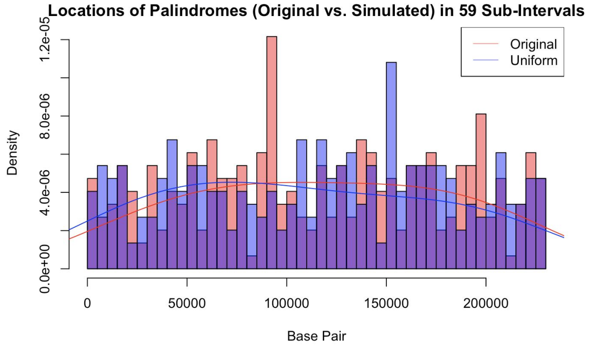


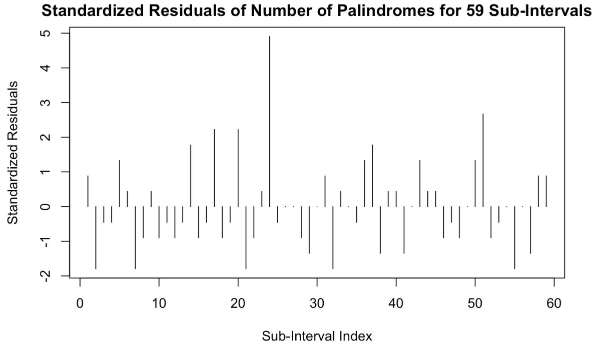

Locations of Palindromes (Original vs. Simulated) in 50 Sub-Intervals


Standardized Residuals of Number of Palindromes for 50 Sub-Intervals


Locations of Palindromes (Original vs. Simulated) in 55 Sub-Intervals


Standardized Residuals of Number of Palindromes for 55 Sub-Intervals


Locations of Palindromes (Original vs. Simulated) in 59 Sub-Intervals


Standardized Residuals of Number of Palindromes for 59 Sub-Intervals

Residuals:

Since most of the p-values for the above testing are small, there is a reason to doubt the fit of the distribution. We then used a residual plot to determine where the lack of fit occurs. Since Most of of the standardized residuals having values that are larger than 3 indicates that a lack of fit. Further, it is apparent that from the residual plots, the big spikes in the middle of the graph corresponds to the spikes discovered in Scenario one.

Goodness of Fit Test Statistic (k represents the number of sub-intervals)

Null Hypothesis (H0): Locations are distributed according to Uniform distribution.
Alternative Hypothesis (H1): Locations are not distributed according to Uniform Distribution.

If we divide the interval into 30, 35, 45, 50, 55, and 59 sub-intervals and examine the frequency of palindromes found in each interval, we would expect the counts of locations found in the sub-intervals follow uniform distribution if our null hypothesis is true. By setting lambda to the expected counts of palindromes distributed uniformly in each interval, we divided n over K. We then calculated the chi-squared statistic and the p-value measuring the level of fitness Uniform distribution to our actual data. The p-values are recorded in *Table* below. At significant level $\alpha = 0.05$, by observing mostly low p-value on cases where the number of sub-intervals are 30, 50, 55, and 59, we conclude that it appears that the Uniform may not be a reasonable initial model for the distribution of locations of our actual data.

Table 1. Chi-Sqr. Table on Goodness of Fit Test for Location of Palindromes in Non-Overlapping Regions

K	30	35	45	50	55	59
X-squared	40.689	32.243	50.318	66.5	70.047	92.682
df	29	34	44	49	54	58
p-value	0.07328	0.5539	0.2376	0.04864	0.06997	0.002578

Inference:

Since most of the p-value of this chi-square goodness of fit test is smaller than 0.05, the chance of observing a test statistic at least as large as ours under the random scatter model is small. We see that deviations as large as ours (or larger) are very unlikely. In addition, having values of the standardized residual larger than 3 suggests that the probability model of a uniform distribution is lack of fit. Hence, we conclude that it appears that uniform is not a reasonable initial model.

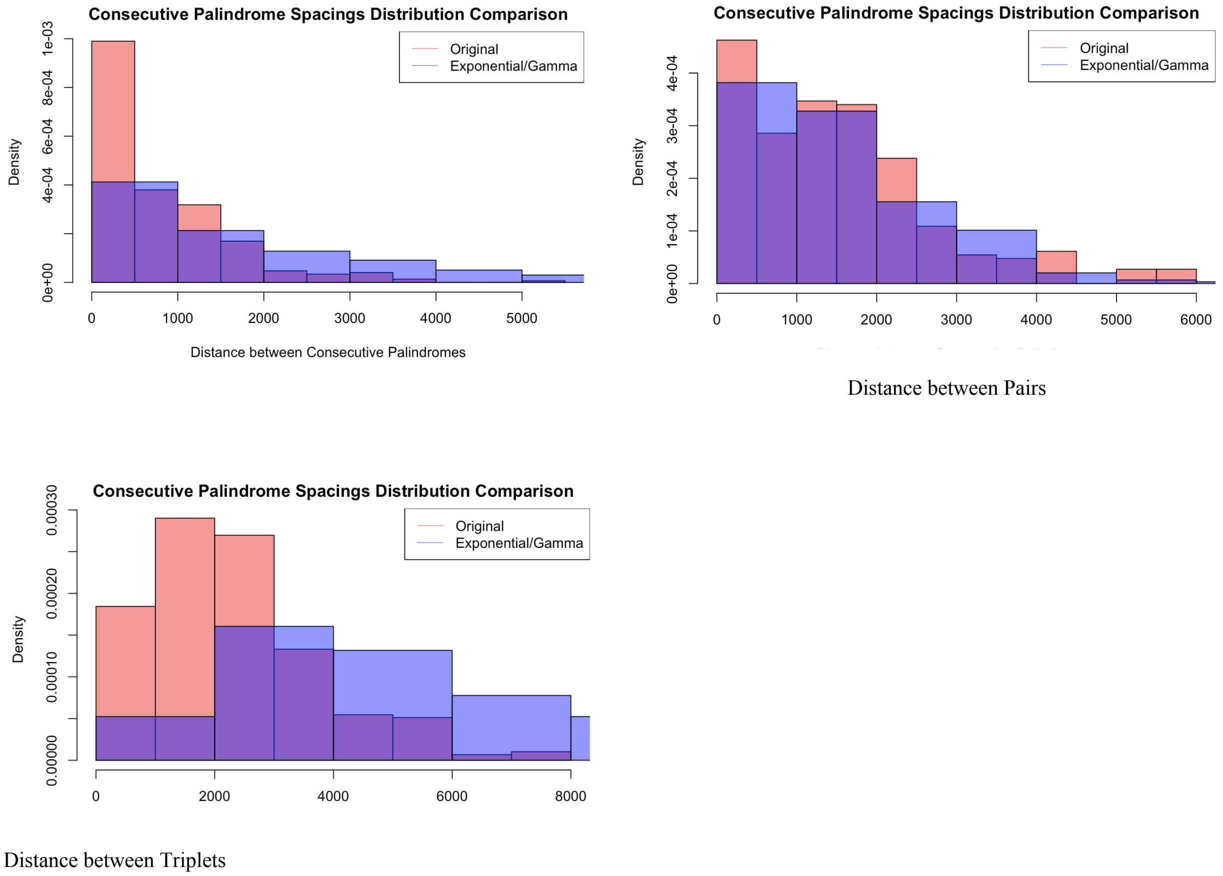
Inference:
Null Hypothesis: Locations are distributed uniformly.
Conclusion: Since p-value of this chi-square test is smaller than 0.05, we reject the null hypothesis. It indicates that deviations as large as ours are not so likely. Hence, we conclude that it appears that uniform is not a reasonable initial model.
Residual Conclusion:
Since the value of the standardized residual is larger than 3, the probability model of a uniform distribution is lack of fit.

For spacings, we investigated the spacings between consecutive palindromes, pairs of palindromes, and triplets of palindromes. The distributions of these spacings are Exponential, Gamma with shape parameter 2, and Gamma with shape parameter 3, all with rate parameter of #palindromes/#base pairs, respectively. We then divided each of these spacings into five categories in terms of length of spacings, then conducted a chi-squared goodness of fit test between the actual spacings and spacings calculated via the respective Exponential and Gamma distributions. The p-values are all below 0.05, meaning that there is a very small probability that the spacing deviations seen in our data from the Exponential and Gamma distributions happened by chance. Therefore, we conclude that the spacing in our data do not follow an Exponential or Gamma distribution, indicating a possibility of clusters.

Spacings

```
Chi-squared test for given probabilities (Exponential)

data:  observed.spacings
X-squared = 13.949, df = 3, p-value = 0.002975

        Chi-squared test for given probabilities (Gamma 2)

data:  observed.spacings
X-squared = 329.54, df = 3, p-value < 2.2e-16

        Chi-squared test for given probabilities (Gamma 3)

data:  observed.spacings
X-squared = 1752.6, df = 3, p-value < 2.2e-16
```

Figure 7. Histograms of distributions of the spacings between the consecutive palindrome



The spacings for the aforementioned singlets, pairs, and triplets of palindromes were visualized. From this we can visualize the departure from the distributions. Therefore, we can suspect a cluster of palindromes.

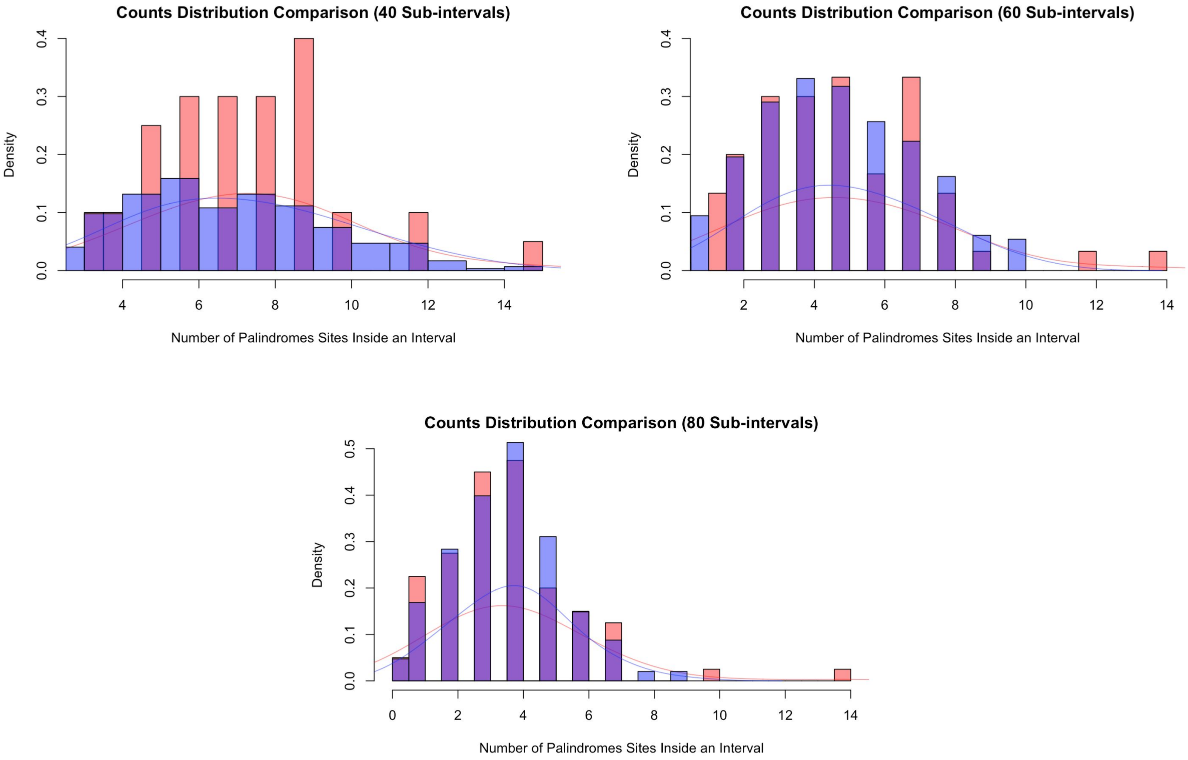
Scenario 3: Counts

The entire dataset is divided into 40, 60 and 80 sub-intervals. We want to investigate how the distribution of the counts of palindromes in each interval fits poisson distribution. In order to achieve our goal, we used graphical methods, goodness of fit testing statistic, and standardized residuals to examine the differences between the two distributions.

Graphical Methods

Figure 8 depicts histograms of the number of palindromes in non-overlapping regions of length 5,733 base pairs, 3,822 base pairs, and 2,867 base pairs. In this visualization, the red graph represents the actual data, and purple represents generated poisson distribution. From visualization, the frequency of counts in each bin varies largely. Again, there appears to be outliers in the distribution's right tail of the original data, whereas outliers are not as apparent in any of the random scatters, indicating the data's departure from a random scatter.

Figure 8. Histogram of Number of Palindromes in Non-Overlapping Regions



Goodness of Fit Test Statistic (k represents the number of sub-intervals)

Null Hypothesis (H0): Counts are distributed according to Poisson distribution.
Alternative Hypothesis (H1): Counts are not distributed according to Poisson Distribution.

If we divide the interval into 40, 60, and 80 sub-intervals and count how many data points are inside each of them, we would expect the counts follow Poisson distribution if our null hypothesis is true. In theory, if our null hypothesis is true, the counts for each method of sub-interval division should follow Poisson processes. By setting lambda to the expected counts of palindromes per interval, we calculated the expected counts for each interval and calculated the chi-squared statistic and the p-value measuring the level of fitness Poisson distribution to our actual data. The p-values are recorded in *Table* below. At significant level $\alpha = 0.05$, by observing high p-value on cases where the number of sub-intervals are 40, 60, and 80, we fail to reject the null hypothesis and conclude that it appears Poisson can be a reasonable initial model.

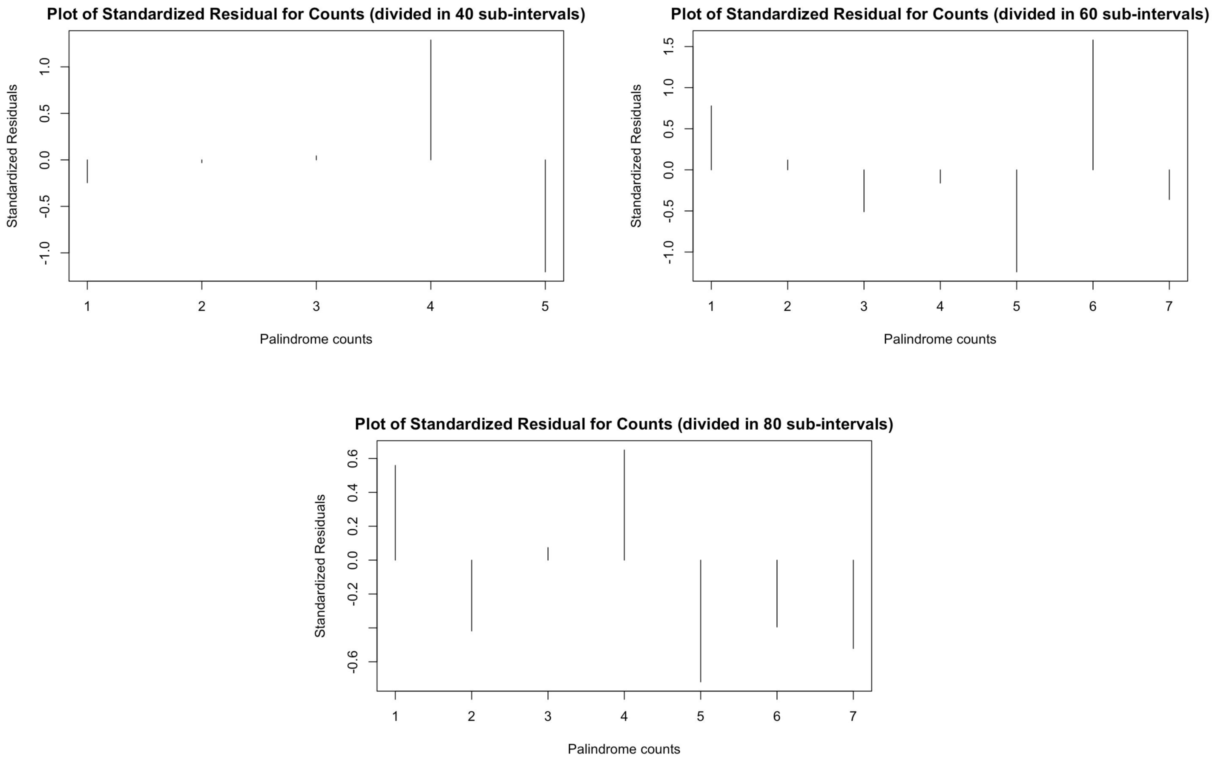
Table 2. P-Value on Goodness of Fit Test for Number of Palindromes in Non-Overlapping Regions

K (number of sub-intervals)	40	60	80
P- Value	0.37	0.41	0.87

Visualizations on Standardized Residuals

We graphed the residuals and found out that in each of the graph, all residuals are less than 3. Therefore, we conclude that it appears that the data fits Poisson distribution.

Figure 9. Residuals of Number of Palindromes in Non-Overlapping Regions



Inference

Since p-value of this chi-square test is greater than 0.05, we will fail to reject the null hypothesis. From this computation, we see that deviations as large as ours are very much likely to be observed. Hence, we conclude that it appears that Poisson is a reasonable initial model for the data.

Scenario 4: Test Statistic for Maximum Cluster

We use $\alpha = 0.05$ as our threshold, the table below shows the probability of the chance that maximum count over m intervals is larger than or equal to k, where k is the maximum counts generated from different length of intervals. *Table 3* shown below has three different amounts of intervals, which are 40, 60 and 80, and each three separation has different lamda and probability. “lamda” shown in the first column is estimated by using the method of MLE since we assume the model follows the Poisson distribution. “Interval_length” is shown in the second column, which is calculated by dividing the corresponding amount of intervals from the total length. The third column is the “p-value”, which the first one is above α and the following two are below. “Maximum counts” are generated by taking the maximum count of palindrome among all the intervals , and “interval” is the specific interval which contains the maximum count. The p-value above α indicates the cluster is not unusual and the replication site is undetected since the regions examined are too large. This explains why from *Table 3*, the probability of getting the maximum cluster larger than k decreases as the length of interval decreases. However, once we narrow the intervals, the cluster is well detected under $\alpha = 0.05$. The p-value (0.002693866) becomes very small when we seperate the total length into 80 intervals, which can be inferred that it is very unlikely for the unusual cluster between the interval of [91700, 94600] to occur by chance. In addition, the interval where the maximum cluster is located for the three different interval_length are all in the range of [91700, 94600], which implies the maximum amount of palindrome gathered within that particular interval is unusual and thus may be a potential replication site.

Table 3: Location of maximum counts correlated with different interval lengths

	lamda <dbl>	**interval_length** <dbl>	**probability** <dbl>	**maximum counts**	**interval**
40	7.400000	5733.850	0.308448064	15	(9.17e+04,9.75e+04]
60	4.933333	3822.567	0.036209375	14	(9.17e+04,9.56e+04]
80	3.700000	2866.925	0.002693866	14	(9.17e+04,9.46e+04]

Scenario 5: Additional Hypothesis

From reading the article, we are interested in finding whether there is a relationship between the distribution of age of HIV positive people and the distribution of age of total population. In order to test that, we first separate people into four age groups: 0-20 is group 1; 21-40 is group 2; 41-60 is group 3; 61+ is group 4. From the data file, since people older than 80 are recorded as 80+, and we only cares about people older than 61 for the last bin, we reassign all the ages of those people to 80. For the column of HIV, we proceed data cleaning by extracting the rows of people that are unknown with their status of HIV. Therefore, the total population decreases from 2174 to 2134. The percentage of people in the age group is calculated by the number of people in the group divided by the population size. The percentage of people with positive HIV in each group is calculated by the number of positive HIV people in that group divided by the total number of HIV positive people in the population. The null hypothesis is H0: The distribution of age of HIV positive people matches the distribution of age of total population, and the alternative hypothesis is H1:The distribution of age of HIV positive people does not match the distribution of age of total population. After calculation, we find that the p-value is equal to 0.92, which is much larger than the significance level 0.05, thus we fail to reject the null hypothesis and therefore conclude that the distribution of age of HIV positive people matches the distribution of age of total population.

Figure10. Standardized Residual plot for Age and HIV Positive

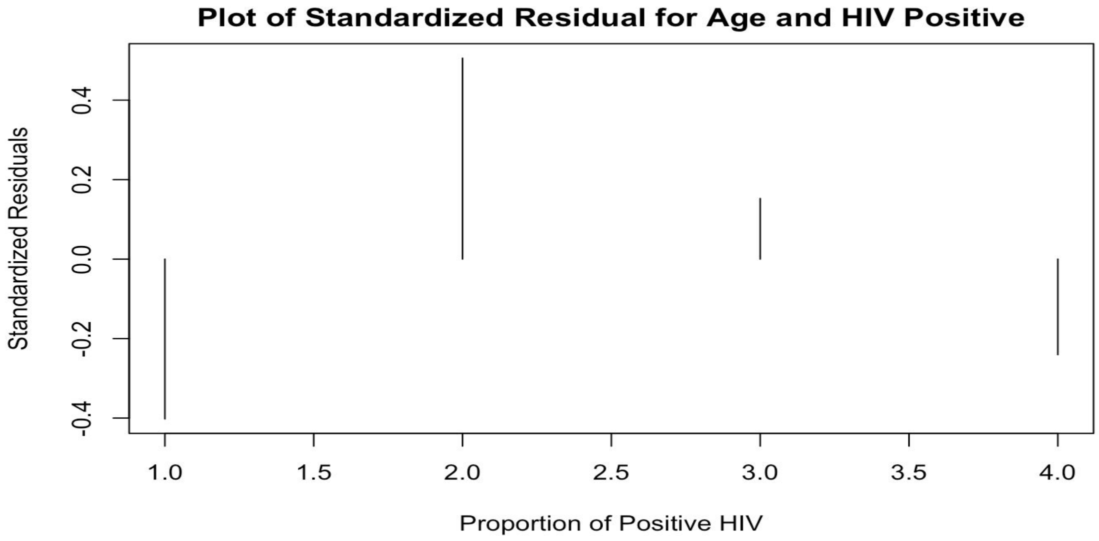


Table 4: Distribution of population between age and likelihood of HIV

Different age group	Percentage of people in that age group/population	Percentage of people that are positive in HIV/Total amount

		of people are positive in HIV
0-20	0.39	0.14
21-40	0.28	0.55
41-60	0.20	0.27
61+	0.13	0.04

Chi-Square Goodness of Fit Test

P-value	0.92

Inference:

Since the P-value is approximately 0.92, which is larger than the critical value α. This indicates that the chance of observing a test statistic at least as large as ours under the distribution of age of total population is large, so it’s very likely that the distribution of age of HIV positive people matches the distribution of age of total population people in different age groups. Therefore, we can conclude that the total population with different ages have the same chance of getting HIV. However, the association or correlation between ages and HIV infection cannot be proved because the situation, in which the distribution of age of HIV positive people matches the distribution of age of total population, could happen by chance. But since the p value here is very large, we can be skeptical about the existence of their association and correlation.

Discussion/Conclusion

From Scenario 1, we visually spotted that there appears to be outliers in the distribution’s right tail of the original data, possibly corresponding to the cluster discussed earlier around the 90000th base pair. The data also appears to deviate from a uniform scatter, hinting at the need for further statistical testing to determine whether or not this indicates a potential replication site.

From Scenario 2 and 3, we used visualizations and goodness of fit statistical testing method to compare our actual data against simulated uniform data. Having small p-values from the goodness of fit testing on locations and spacings indicates that uniform distribution may be a lack of fit to our data. The graphs of residuals further supported the idea that locations may not be distributed uniformly, and there might be big clusters interrupting the uniformity of the data distribution. In addition, we have seen that the spacings between consecutive palindromes, pairs of palindromes, and triplets of palindromes deviate from their respective exponential/gamma distributions.

From Scenario 4, after dividing the total length into different interval lengths, we calculated the probability of the chance that maximum count over m intervals is larger than or equal to the maximum counts generated from different length of intervals. Particularly, the p value of the maximum counts of the palindromes occurring in the interval between 91700 and 94600 is very small, which indicates that the clusters occurring in that interval is not possibly by chance. Additionally, the intervals of the clusters calculated from the three different amounts of intervals all include the the interval between 91700 and 94600, so this interval with the greatest amount of palindromes indicate a potential origin of replication. Therefore, biologists can focus on the interval between 91700 and 94000 to find the origin of replication efficiently.

From Scenario 5, we use chi-square Goodness of Fit Test to test whether the distribution of age of HIV positive people matches the distribution of age of total population, and p-value of 0.92 is generated. Thus, under significance level $\alpha = 0.05$, we accept the null hypothesis and conclude it’s very likely that the distribution of age of HIV positive people matches the distribution of age of total population people in different age groups, which indicates that the total population with different ages have the same chance of getting HIV. However, the association or correlation between ages and HIV infection cannot be proved because the situation, in which the distribution of age of HIV positive people matches the distribution of age of total population, could happen by chance.

All of these scenarios have shown that our data does not follow the uniform distribution, via formal statistics as well as visualization and simulations. Therefore, we highly suspect the existence of clusters and palindromes, and thus potential replication sites.

Overall, in order to find the origin of replication, we suggest the biologists cut DNA into segments and test for the unusual big cluster. Since the process of detecting viruses is expensive and time consuming, it is helpful for the biologists looking for the unusual cluster before actually conducting the experiments in order to reduce the budgets. Since the origin of replication for two viruses from CMV family is marked by complimentary palindromes, biologists can focus on the certain segment of DNA (for example the interval between 91700 and 94000 pairs in this case) and save the amount of testings need to be conducted to identify the location of unusual cluster of complementary palindromes and further investigate the origin of replication.

Theory

Goal

Our goal is to understand the statistics model that describes "counts" of the number of palindromes and "uniformity" of random distribution of palindromes. We want to determine the estimation procedure in the model. Also, we want to find statistical discrepancies between a model with clusters and model without clusters. The questions we want to answer include whether the model is a good model, what is hypothesis test and how is uniform distribution related to our problem.

The Homogeneous Poisson Process

The homogeneous Poisson process is a process that arises naturally from the notion of points randomly distributed on a line without obvious regularity. Many random phenomena can be modeled by the homogeneous Poisson process such like arrival times of calls at an exchange, the decay times of radioactive particles and positions of stars in parts of the sky. The homogeneous Poisson process has three major characteristics: 1. The rate λ at which points occur does not change with location. 2. The number of points falling in different regions are independent. 3. No two points can hit at the same place.

Counts of the number of points in different regions follow Poisson distribution with rate λ. P(k points in a unit interval)= $\frac{\lambda^k * e^{-\lambda}}{k!}$ and the expected number of hits per unit interval is λ. Since in most cases λ is unknown, there are two methods of estimation: the method of moments uses the empirical average number of hits per unit interval as an estimate, and the other one is maximum likelihood method. These two methods result in the same estimator in Poisson distribution.

Checking the Homogeneous Poisson Process

The Poisson process is a good reference model for making comparison because it is a natural model for uniform random scatter. We want to show that our case can be modeled by homogeneous Poisson process. Firstly, the strand of the DNA is the line in the model, and the location of a palindrome is a point on the line. According to uniform random scatter model, palindromes are scattered randomly and uniformly across the DNA. Also, three properties of homogeneous Poisson process are satisfied: the chance that one piece of DNA has palindrome in it is the same for any piece of DNA (property1). The number of palindromes in any piece of DNA is independent of the number in another without overlapping (properties 2&3)

Chi-Square Goodness of Fit Test

Chi-Square Goodness of Fit Test is applied when we have a categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution. Sometimes a parameter of the distribution needs to be estimated in order to compute the probabilities. In this case, data are used to estimate the

unknown parameter(s). The measure of discrepancy between the sample counts and the expected counts is $\sum_{j=1}^{m}(jth\ observed\ count - jth\ expected\ count)^2/jth\ expected\ count = \sum_{j=1}^{m}(\ Nj\ \text{-}\ \mu j)^2/\mu j$.

To compute p-value we use χ^2 distribution. If the probability model is correct, then the test statistic has approximate chi-squared distribution with m-k-1 degrees of freedom, where m is the number of categories and k is the number of parameters estimated to obtain the expected counts. χ^2 with degrees of freedom m-k-1 is a continuous distribution on the positive real line and the density has a long right tail. As the degrees of freedom increase it starts to look symmetric and a lot like normal. If the p-value is small, then there is a reason to doubt the fit of the distribution. If the test statistic is greater than χ^2 with degrees of freedom m-k-1 and significance level α, the null hypothesis will be rejected. If the test statistic is less than χ^2 with degrees of freedom m-k-1 and significance level α, the null hypothesis will not be rejected.

Locations and the Uniform Distribution

Under the Poisson process model for random scatter, if the total number of hits in an interval is known, then the positions of the hits are uniformly

scattered across the interval. The Poisson process on a region can be seen as a process

that first generates a random number, which is the number of hits, and then generated locations for the hits according to the uniform distribution.

Exponential and Gamma Distribution

In the Poisson process, distances between successive hits follows an Exponential distribution. P (the distance between the first and second hits > t)=

P (no hits in an interval of length t) = $e^{-\lambda t}$. Distances between the hits that are two apparatus, follows a Gamma distribution with parameters 2, λ.

Suppose (N(t), t$\geq$0) is a Poisson process with rate λ, N(s+t) - N(s)~Poisson(λt). Then, time of nth event: Sn = min{t : N(t) = n}, Sn~Gamma(n, λ). Time between events: T1 = S1, Tn = Sn - Sn-1 for n$\geq$2, then Tn~Exponential(λ).

Maximum Numbers of Hits

Under the Poisson process, the numbers of hits in a set of non-overlapping and equal intervals are independent observations from a Poisson distribution. This means that the maximum number of hits in a set of intervals behaves the same as the maximum of independent Poisson random variables. Let m be the total number of intervals then

P(maximum count over m intervals $\geq$k)
= 1-P(maximum count over m intervals $\leq$k)
=1-P(all interval counts<k)
=1-$P(frst\ interval\ counts < k)^m$
= 1- $[\lambda^0 e^{-\lambda} + ... + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^m$

Therefore for a given estimate of λ, from the above expression the approximate chance that the greatest number of hits is at least k can be find. If this chance is unusually small, then it provides evidence for a cluster that is larger than the expected from the Poisson process. The maximum palindrome counts can be used as a test statistic, and the computation above provides the p-value for the test statistic.

Method of estimation

Let X1,...Xn be iid, from a Poisson distribution with unknown rate parameter λ. MOM is one estimation technique that proceed as follows:

1. Find E(X) where X has Poisson distribution with rate λ
2. Express λ in terms of E(x)
3. Replace E(X) with $\bar{x}$ to produce an estimate of λ, called $\widehat{\lambda}$.

For Poisson distribution, E(x) = $\lambda \rightarrow \bar{x} = \widehat{\lambda}$ If higher moments need to be computed then E(x^2) is replaced with $\sum_{i=0}^{n} xi^2/n$

- Maximum Likelihood Estimator

Let X1,...Xn be iid from a Poisson distribution with unknown rate parameter λ. Maximum likelihood method searches among all Poisson distributions to find the one that places the highest chance on the observed data. For Poisson distribution, the chance of observing x1,xn is $\frac{\lambda^{x1}}{x1!} e^{-\lambda} * ... \frac{\lambda^{xn}}{xn!} e^{-\lambda} = \frac{\lambda \sum_{i=0}^{n} xi}{\prod_{i=0}^{n} xi!} e^{-\lambda} = L(\lambda)$ For given data, this is a function of λ that is called the likelihood function. Maximum likelihood estimates the unknown parameter by the $\lambda - value$ that maximized the likelihood function. Since the function is monotonically increasing, the log likelihood function, denoted with l, is maximized at the same values as L. To find the maximum we consider solving the first-order equation

$$\frac{\partial}{\partial\lambda}\ l(\lambda) = \frac{\partial}{\partial\lambda} [\sum_{i} xilog(\lambda) - n\lambda - \sum_{i} log(xi!)] = \sum_{i=1}^{n} /\lambda - n = 0.$$

By solving the last equation for λ we obtain:
$\widehat{\lambda} = \bar{x}$. Maximum-likelihood for continuous distributions is the same. Suppose we have an independent sample x1,..,xn. From an Exponential distribution with the unknown parameter θ.

Now, the Likelihood function, given the data is L(λ)= $\theta^n e^{-\theta \sum_{i=0} xi}$, and the log-likelihood function

l(θ) = nlog(θ)- θ $\sum_{i=0} xi$. By solving the last equation for $\widehat{\theta} = \frac{1}{\bar{x}}$.

Properties of Parameter Estimates

We use mean squared error to compare and evaluate parameter estimates. It is defined as, MSE($\widehat{\lambda}$) = E($\widehat{\lambda}$ $-$ λ)^2 = Var($\widehat{\lambda}$)(variance) + (E($\widehat{\lambda}$) $-$ λ)^2 (squared BIAS). Many of the estimators we use are UNBIASED, but sometimes an estimator with a small bias will have a small MSE. Under certain regularity conditions, as the sample size increases, the Maximum-likelihood estimator, λˆ satisfies λˆ → λ and λˆ ~ N(λ, 1/nI(λ)) where I(λ) is called the Fisher's Information Matrix. Fisher's Information matrix is defined as
$I(\lambda) \ = \ E(\frac{\partial}{\partial\lambda} log\, f_{\lambda}(X))^2 \ = \ -E(\frac{\partial^2}{\partial\lambda^2} \, log\, f_{\lambda}(X))$. Therefore, as n increases, $\sqrt{nI(\lambda)}$ $(\widehat{\lambda} - \lambda) \sim N(0, 1)$. The approximate normal distribution can be used to build the 95% confidence interval for the unknown λ as $\widehat{\lambda} \pm 1.96\sqrt{nI(\lambda)}$.

Hypothesis Tests

The χ 2 goodness-of-fit test and the test for the maximum number of palindromes in an interval, are two examples of hypothesis tests. In hypothesis testing, we assume that the H_0 is true and find out how likely our data are under this model. We can use it to test whether the two distributions have the same average. For example, $\overline{X}$ and $\overline{Y}$ are good estimators of μ1 and μ2. They are independent and asymptotically normally distributed. Since $\overline{X}$ and $\overline{Y}$ has approximately normal distribution, a good candidate for the test statistic is its rescaled version, that under the null satisfies Z = $\frac{\overline{X} - \overline{Y}}{\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}} \sim N(0, 1)$. We call this test statistic a Z-test, as it is based on normal approximations. Then, the p-value is computed as P(|Z| > Z observed). As p-value > 5%, we conclude that the observations support the Null hypothesis. If the p-value < 5%, we would have concluded that the observations do not support the null, and we would reject the null in favor of the alternative.The cutoff of 5% is called significance level of a test. However, the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not. When we reject the null hypothesis, we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision. There are two types of errors: Type I and Type II errors. Type I error = α or the significance of the test, while Type II error := β. Power of a test := 1 − β. Typically α is set in advance and β is computed for various values of the alternative hypothesis. High power is a sign of a good test.

Works Cited

Bradic, Jelena. “Chapter 4: Patterns in Data.” MATH 189 Lecture, UC San Diego, 1 May 2018. Lecture.

Centers for Disease Control and Prevention. "Cytomegalovirus and Congenital CMV Infection." *Saving Lives, Protecting People*, 5 Dec. 2017, www.cdc.gov/cmv/index.html. Accessed 9 May 2018.

Editors of Encyclopædia Britannica. "DNA." *Encyclopædia Britannica*, 6 Feb. 2018, www.britannica.com/science/DNA. Accessed 9 May 2018.

Mandal, Ananya. "What Is Virus." *News: Medical and Life Science*, 3 Aug. 2017, www.news-medical.net/health/What-is-a-Virus.aspx. Accessed 9 May 2018.

"What Is DNA." *U.S National Library of Medicine*, 21 Apr. 2005, ghr.nlm.nih.gov/primer/basics/dna. Accessed 9 May 2018.