Math 189 - Case Study #3

Patterns in DNA

February 28th, 2019

Name	PID
Erin Werner	A12612584
Emma Choi	A12635909
Talal Alqadi	A13816618
Ella Lucas	A13557332
Samantha De La Torre	A13300273

1. Introduction

Human cytomegalovirus (CMV) is a potentially life threatening disease for people with suppressed or deficient immune systems. In order to combat this disease, scientists study the way the virus replicates. Specifically, they study a region of the virus's DNA called the origin of replication, which contains special instructions on how to replicate.

DNA is constructed of a combination of A's, T's, G's, and C's, where A's and T's form a complementary pair and G's and C's form a complementary pair. The DNA code provides the virus with the information to grow, survive, and replicate. However, with only 4 letters, it requires additional ways to encode information. Patterns in the sequence can signal important sites on the DNA, such as the origin of replication. One type of pattern used to highlight an important site is a complementary palindrome, which consists of an order of letters, that when read in reverse, forms the complement of the forward sequence.

CMV is a member of the Herpes Family, along with Herpes simplex and Epstein-Barr Virus. Both Herpes simplex and Epstein-Barr contain complementary palindromes to mark their origin of replication. However, Herpes simplex has a long palindrome, whereas Epstein-Barr has a series of short palindrome sequences. It has been suggested that a cluster of palindrome sequences in CMV play a similar role to the clusters in Epstein-Barr or the long palindrome in Herpes simplex and is marking the origin of replication. In order to identify the location of the origin of replication, DNA is cut into short sequences, and tested to see if it can replicate. If it can still replicate, then the origin of replication is contained in that sequence. However, this has proved to be a time consuming process. The goal of this lab is to search for unusual clusters of complementary palindromes in CMV because they could signal the location of the origin of replication.

2. Data

The DNA sequence of CMV was published in 1990 (Chee et al.). Search algorithms that screened the DNA sequence for different types of patterns were developed by Leung et al in 1991. As a result, 296 palindromes were found, and they all were at least 10 letters long. The longest palindromes found were 18 letters long and occurred in locations 14719, 75812, 90763 and 173893 along the sequence. In this dataset, palindromes shorter than 10 letters were ignored. The CMV DNA is 229,354 letters long. The data is not independent, identically distributed (iid) due to the fact that if pairs are closer together, they are more dependent compared to ones father away from each other.

In order to start grouping the data that consists of the 296 palindromes found, we can segment the DNA chain into intervals of base pairs. Then, we can compute the number of palindromes found within each of the subintervals. This process is demonstrated in Figure 2.1, with an interval size of 4,000. From this histogram, it is fairly easy to see that despite the length of the interval, clusters of palindromes appear in at least two locations, those being the 93,000th and the 195,000th pair of DNA. This is enough evidence to formulate a hypothesis that claims that the clusters at theses two locations are exceptions within the typical structure of the DNA chain. This would then indicate that the clusters are not due to chance.

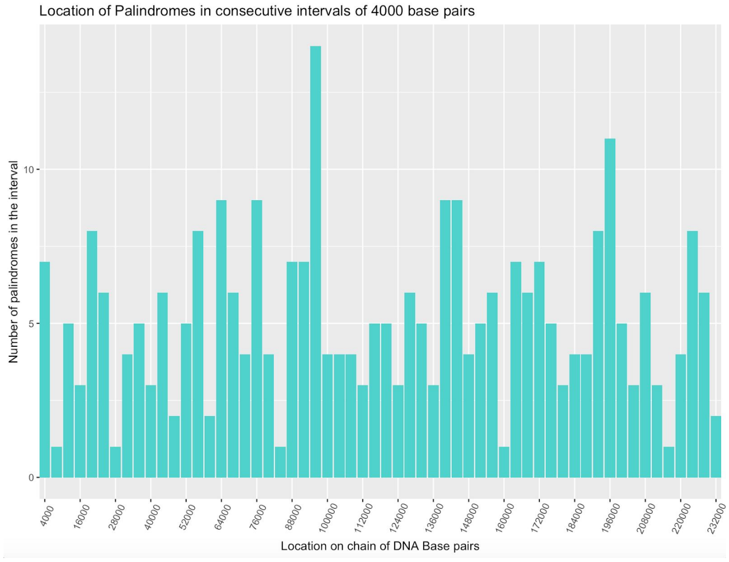


Figure 2.1. Histogram of palindromes in a subinterval.

To help strengthen this hypothesis, we ran a simulation with 296 random samples from the same range of values. We then formed a similar histogram, but with the randomly generated data. By comparing the histogram of the actual palindromes to the histogram based on randomly generated numbers, like Figure 2.2, we can see that the random sets of numbers present no pattern of clusters at any given point, no matter what size intervals we use.

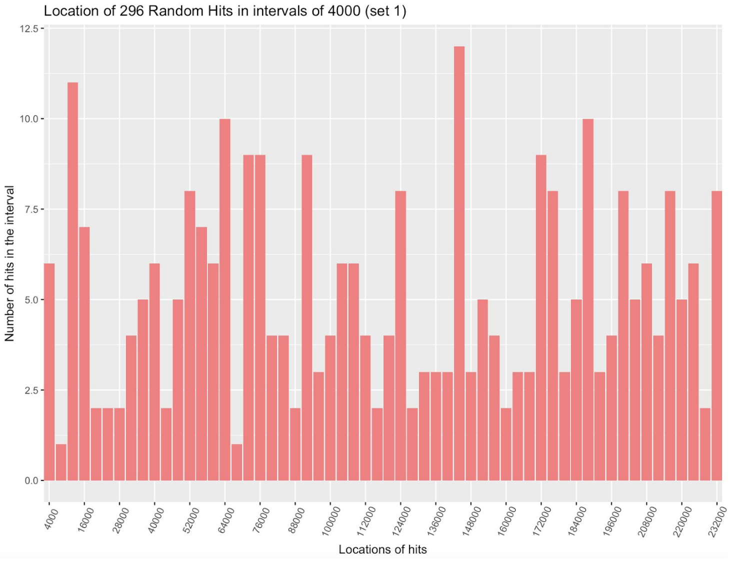


Figure 2.2. Histogram of random scatter baseline simulation.

Another useful observation is the number of intervals in relation to the palindromes. The resulting histogram, Figure 2.3, reveals an exponential distribution of the actual data. We can also see that the observed palindromes present higher spikes of numbers of palindromes per intervals. No matter the length of the intervals, there always seems to be one or two outliers of intervals containing a higher number of palindromes.

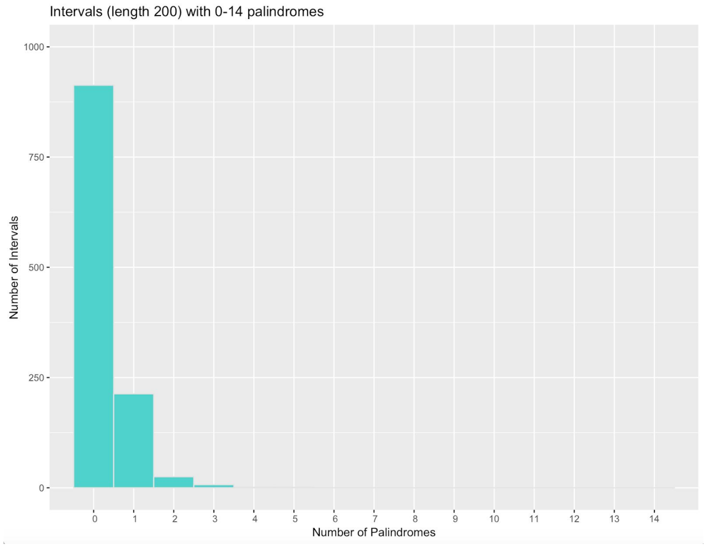


Figure 2.3. Histogram of palindrome count conditional to number of intervals.

However, the same result is not produced with the randomly generated data. We can observe that the intervals of the random hits do not display the same such outliers. In addition, there does not appear to be any consistent pattern of clusters of hits with the random numbers. Therefore it would seem logical to deduce that the outliers on the DNA are atypical and worth examining for the replication code.

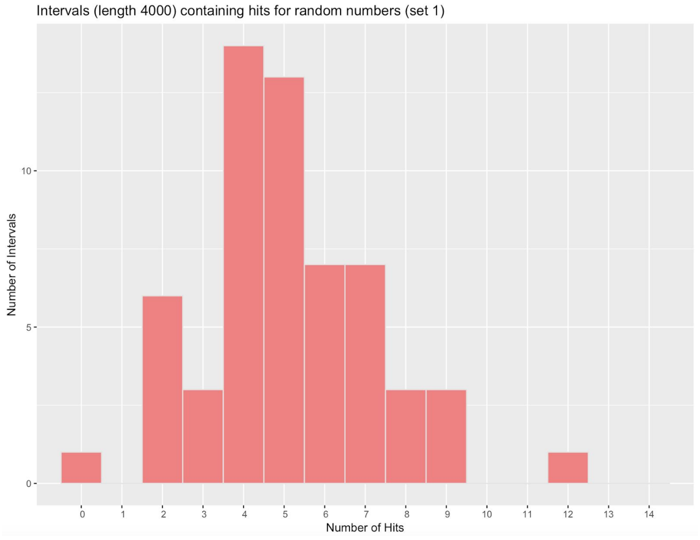


Figure 2.4. Histogram of “hits” count conditional to number of intervals from sample.

We can also analyze the spaces between palindromes as a way to help determine cluster locations and significance. There will most likely be clustering if the spaces between palindromes are small.

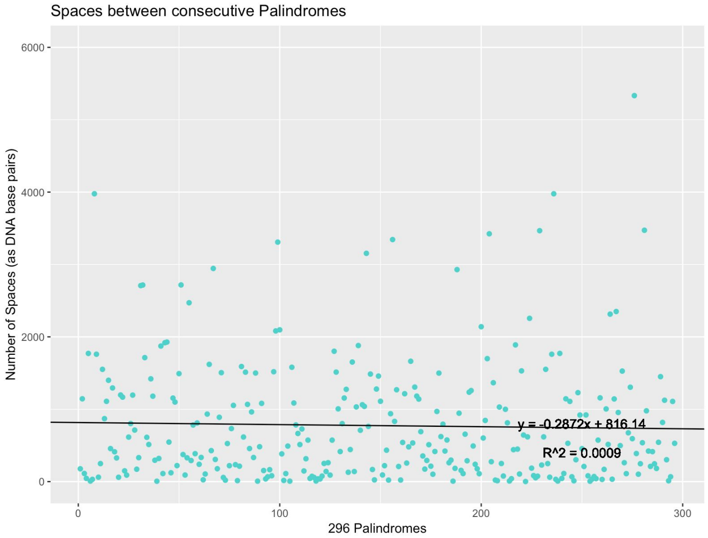


Figure 2.5. Scatter plot of spaces between palindromes.

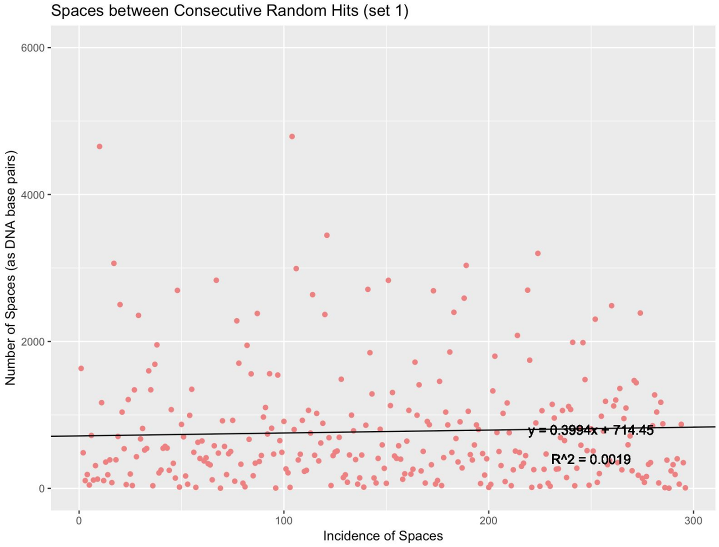


Figure 2.6. Scatter plot of spaces between random hits.

Yet, the results, shown in Figure 2.5 and Figure 2.6, don’t produce any patterns that help us to draw any conclusions. Thus, the scatter plots are not very useful for determining abnormalities. So, other techniques will be explored throughout the report in order to better determine cluster locations and significance.

3. Background

In 1944, Avery, MacLeod, and McCarty showed that DNA was the carrier of hereditary information. Nine years later, Watson and Crick proved that DNA has a double helical structure composed of two long chains of nucleotides. Each nucleotide has a deoxyribose sugar, a phosphate, and a base. There are four bases, adenine (A), thymine (T), guanine (G), and cytosine (C). The two strands of DNA are connected by the bases and each base has to be bonded to its complementary pair. Adenine is complementary to thymine, thymine to adenine, guanine to cytosine, and cytosine to guanine. The DNA in CMV contains 229,354 complementary base pairs, whereas the DNA in humans contains more than 3 billion base pairs.

Viruses are made up of two main parts: a DNA molecule and a protein shell called a capsid that encloses the DNA. The DNA contains all the necessary information for the virus to survive. Typically, virus DNA is several thousand base pairs in length. CMV is a member of the Herpes family. The incidence of CMV varies geographically by 30-80%. It mostly affects children under the age of five and young adults. It has symptoms similar to that of mononucleosis. The virus stays dormant unless it can reach a productive cycle in which it can reproduce thousands of copies rapidly. Once able to reproduce rapidly, it becomes harmful, particularly to people in immune-depressed states, such as transplant or AIDs patients. By locating the origin of replication, a virologist would have an increased chance of developing a vaccine.

4. Investigation

a. Random Scatter

In order to determine if clusters of palindromes were of chance occurrence or could be a potential replication site we simulated a uniform random scatter, seen in Figure 4.1.1, to compare the CMV DNA Data to. To mirror the CMV data, the random scatter was created using 296 palindrome sites and 229,354 bases that we binned into 42 subintervals.

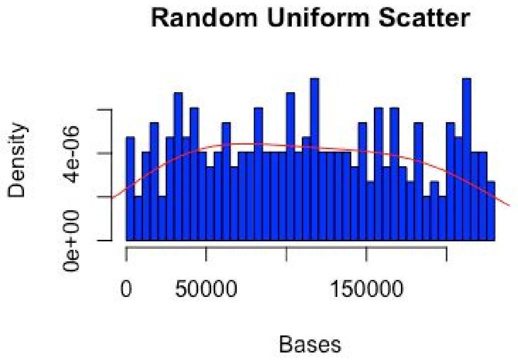


Figure 4.1.1. Random uniform scatter.

Then using the CMV data we created a histogram plotting the counts of palindromes in each subinterval of the DNA. This can be seen in light purple on Figure 4.1.2. The simulated data from the scatter in Figure 4.1.1 was used to create a Poisson Distribution, seen below in light pink, to compare the CMV data to. Based on the overlapping points from the CMV data and the simulated Poisson data, seen below in dark pink, we determined that the CMV data follows a Poisson Distribution.

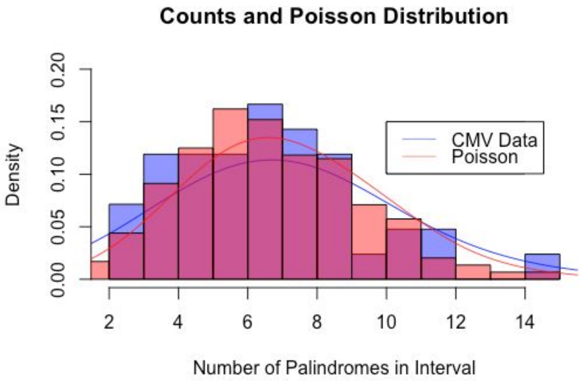


Figure 4.1.2. Histogram of CMV palindrome counts and the poisson distribution.

Additionally, we used strip plots, seen in Figure 4.1.3 and Figure 4.1.4, to compare the palindrome locations of the CMV data and the randomly simulated data. However, they looked fairly similar and did not provide a substantial amount of information into the differences between the two distributions. Therefore, we decided that other methods of comparison, such as the histograms above, were more informative about the data.

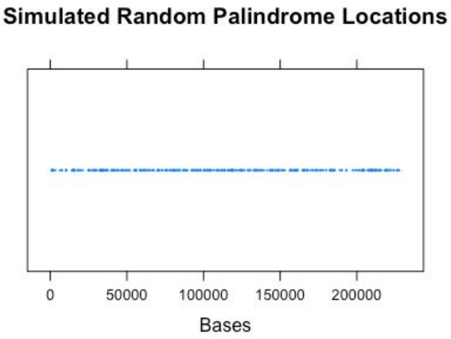


Figure 4.1.3. Strip Plot of Random Simulation Data

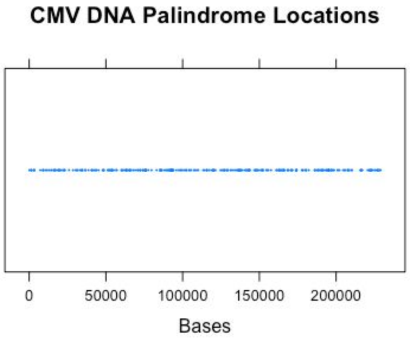


Figure 4.1.4. Strip Plot of CMV DNA Data

b. Locations and Spacings

i. Locations

The location of palindromes important to analyze as it is useful in determining if an abnormally large number of palindromes are found in a certain location of the DNA sequence. This would indicate that the replication site is in that area. To best deduce whether the results are significant, the locations of palindromes will be compared to randomly generated data. The random data is of uniform scatter and span the same range as the CMV data. In order to visualize the locations of palindromes, the DNA sequence was divided into 23 intervals, which is approximately 10,000 palindrome locations per interval. The count of locations within each subinterval is then plotted as a histogram in Figure 4.2.1.

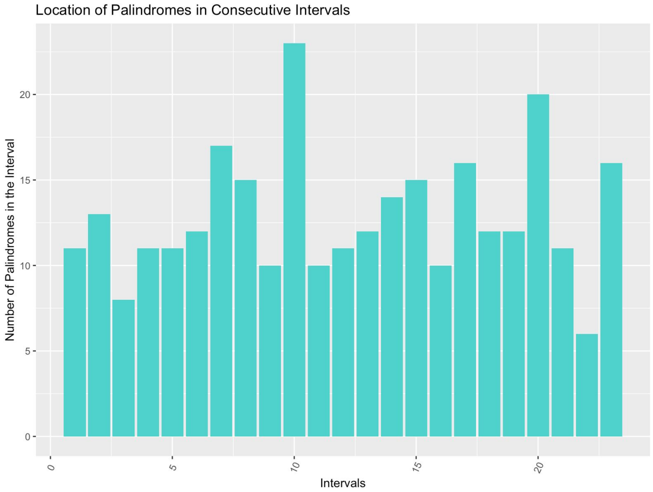


Figure 4.2.1. Histogram of palindromes per interval.

Intervals 10 and 20 contain an abnormally large number of palindromes, but it is possible that the peaks are due to randomness and may not actually be significant.

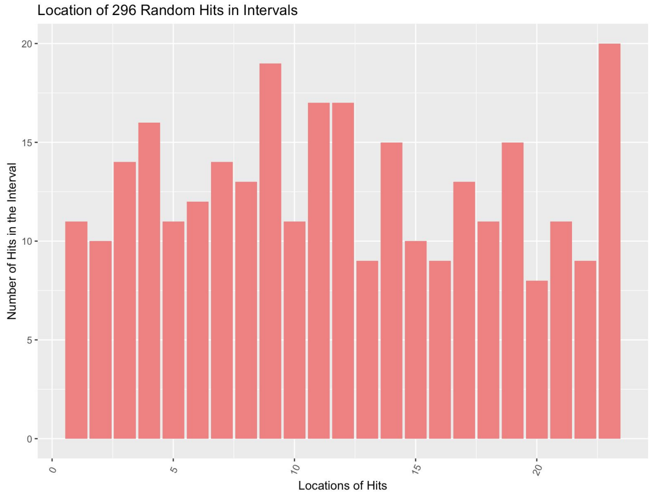


Figure 4.2.2. Histogram of randomly generated data per interval.

In order to determine how significant our results are, we created a similar histogram from the randomly generated data, shown in Figure 4.2.2. This plot helps us figure out if the peaks found in the sample are actually significant by comparison. The random uniform distribution is created by placing 296 palindromes randomly and uniformly across 229,354 locations and dividing the result into the same 23 intervals as done above. The results from the random distribution show a few abnormal peaks. Yet, the peaks are not as extreme as the peaks from the DNA sequence.

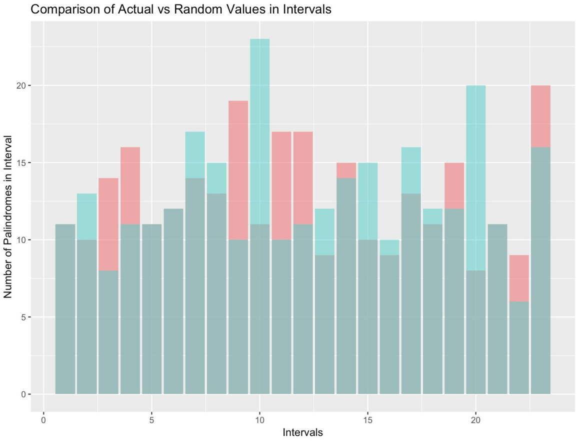


Figure 4.2.3. Histogram of both palindromes and randomly generated data per interval.

To get a more direct comparison, the two graphs are placed over each other. Thus, it is clear that both the sample and the random uniform distribution have large peaks in different locations, but the peaks at intervals 10 and 20 from the DNA sample are more significant than any peak in the random uniform distribution.

To mathematically determine if these abnormally large peaks are actually significant, a chi-squared test is performed. We will assume the null hypothesis, that there are no differences between the two distributions, and if our results are lower than the set tolerance ($\alpha = 0.05$), we will consider the alternate hypothesis, that there actually is a difference between the two distributions. The resulting p-value from the chi-squared test is 0.202. This value is much larger than the 0.05 threshold needed to reject the null hypothesis, so it cannot be concluded that the distributions are different from this test.

Therefore, it is fair to deduce that the peaks in the locations of palindromes from our sample DNA data are likely due to randomness. So, the locations, consequently, do not provide significant evidence that the replication site may be in those locations.

ii. Spacings

The spaces between each consecutive palindrome are also an important feature that can help to determine whether a cluster of palindromes is a potential replication site. As derived from the Poisson process, the distances between two palindromes should follow an exponential distribution. This is best exemplified by plotting a histogram of all the calculated distances between pairs of palindromes from the DNA sequence. To simulate an expected exponential distribution, we used the MLE method with rate λ=0.001. We plotted the two distributions together in order to determine if the palindrome spacing actually follows an exponential distribution, as demonstrated in Figure 4.2.4.

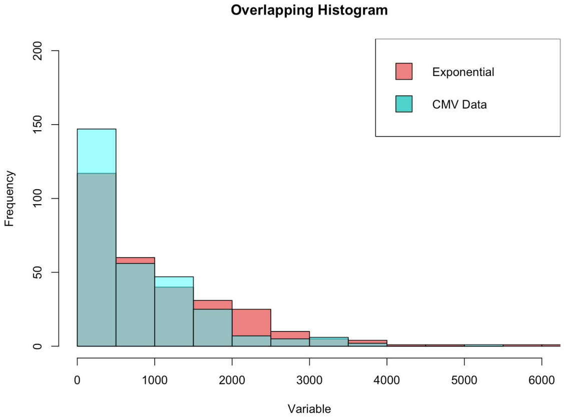


Figure 4.2.4. Histogram of the distribution of spacing between consecutive palindromes compared with an exponential distribution.

To evaluate whether or not the distribution of the DNA palindrome spacings differs significantly from the simulated exponential distribution, we can implement a chi-squared test. The resulting p-value is 0.254, which is still larger than our tolerance of 5%. As a result, we cannot reject the null hypothesis claiming that there are no differences between the distributions of spacing between pairs of palindromes and the exponential distribution.

Furthermore, to provide more evidence for this, we can calculate the double spacing between palindromes. This double space is the separation between the 1st and 3rd palindrome, the 2nd and 4th palindrome, and so on for the rest of the DNA sequence. The histogram of the calculated double spaces and the comparative exponential distribution are shown in 4.2.5.

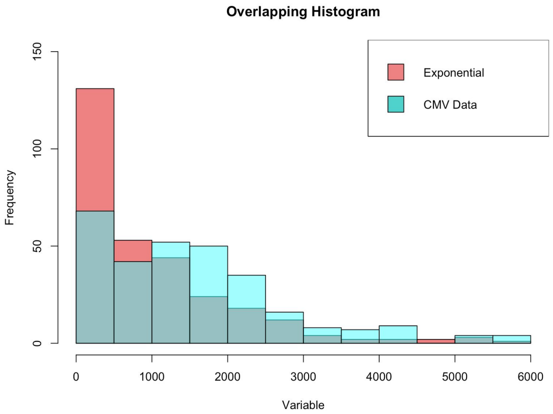


Figure 4.2.5. Histogram of the distribution of double spacing between every other palindrome compared with an exponential distribution.

Once again, we perform a chi-squared test and get a p-value of 0.248. So, we do not reject the null hypothesis claiming that there is no difference between the distribution of double spacing between palindromes in our CMV data and an exact exponential distribution. Thus, it is fair to conclude that the spacing between pairs of palindromes follow an exponential distribution.

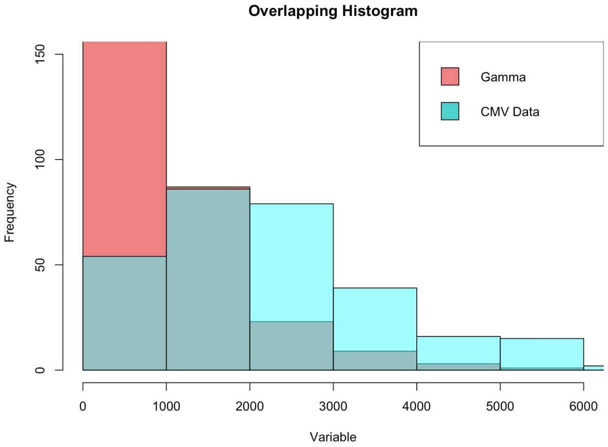


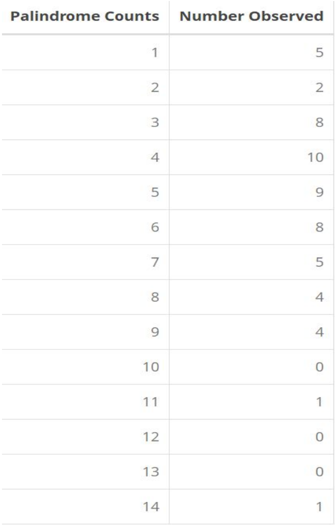
Figure 4.2.6. Histogram of the distribution of triple spacing between every third palindrome compared with a gamma distribution.

As there is substantial evidence that the single and double spacing of palindrome pairs follows an exponential distribution, it makes sense that other properties of the homogeneous Poisson process would apply to our CMV data. For instance, the triple spacing of palindromes should follow a gamma distribution. Triple spacing reflects the calculated distance between the 1st and 4th palindrome, the 2nd and 5th palindrome, and so on for the rest of the DNA sequence. The resulting histogram of triple spacing and a comparative gamma distribution are shown in Figure 4.2.6. The plot demonstrates that the two distributions are relatively similar. To further confirm this property, we performed a chi-square GOF test and obtained a p-value of 0.2484.

Once again, we do not reject the null hypothesis claiming that there is no difference between the distribution of triple spacing between pairs of palindromes in our DNA sequence and an exact gamma distribution.

c. Counts of Palindromes

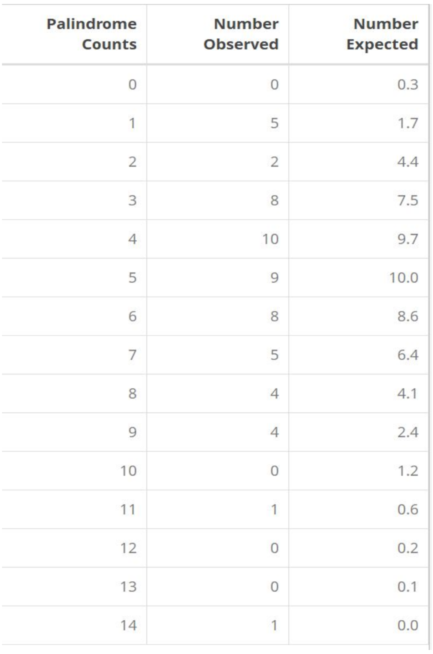
To calculate the counts of palindromes, we first begin by choosing an interval size of 4000. This interval size allows us to have many zero hits, along with not having too big of hits. By dividing the interval size with the total number of complementary pairs of letters in reference to the cmv DNA, which equates to $\frac{229,354}{4000} \simeq 57\ \mathit{intervals}$. Since the lambda is missing in the poisson distribution, we assume lamba to be the hits per interval, so: $\lambda = \frac{296}{57} = 5.16$.



Palindrome Counts	Number Observed
1	5
2	2
3	8
4	10
5	9
6	8
7	5
8	4
9	4
10	0
11	1
12	0
13	0
14	1

Table 4.3.1. Distribution of palindrome counts with number of hits observed.

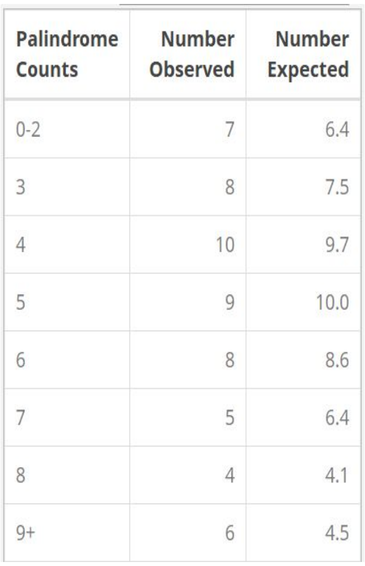
Table 4.3.1 represents the distribution of the number of palindrome hits within each 4000 interval, with the actual numbers observed. From first glance, the interval 4000 seems reasonable with minimal zero observed hits (only 2). The maximum number of palindrome hits in an interval is 14, while the lowest was 1. The maximum number of observed hits is 10 with 4 palindromes within a 4000 interval. In this case, we assume the null hypothesis to be the data follows a poisson distribution. Using the Chi-Squared test statistics, we confirm that the poisson model is applicable, and that the test statistic distribution has an approximate Chi-Squared distribution, as the sum in the Chi-Squared test is 1.0494 with 6 degrees of freedom. Using the information above, and the predicted rate (λ), we calculate the estimate values using the poisson distribution equation.



Palindrome Counts	Number Observed	Number Expected
0	0	0.3
1	5	1.7
2	2	4.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.4
8	4	4.1
9	4	2.4
10	0	1.2
11	1	0.6
12	0	0.2
13	0	0.1
14	1	0.0

Table 4.3.2. Distribution and expected value of counts.

Table 4.3.2 allows us to see that the expected values are not that far off from the observed values, with the highest expected value being 10, which is true.



Palindrome Counts	Number Observed	Number Expected
0-2	7	6.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.4
8	4	4.1
9+	6	4.5

Table 4.3.3. Grouped distribution and expected values with interval size 4000.

In Table 4.3.3, we bring the range of expected and observed values closer to each other, we combine and collapse the 0-2 and 9+ palindromes counts, allowing us to get rid of the zeros, and have all values be at least 4. Furthermore, comparing the observed data to the expected data, the chi squared goodness of fit test allows us to compute the chance of observing a test statistic that follows a chi squared distribution. With a significant level of 0.05 and degrees of freedom 8

- 2, the p value with the chi squared test is 0.983. In addition, the chi squared cut off value is 12.59, which is much higher than the observed test statistic of 1.04. Thus, the chi squared test allows us to accept the null hypothesis that the Poisson is a reasonable model.

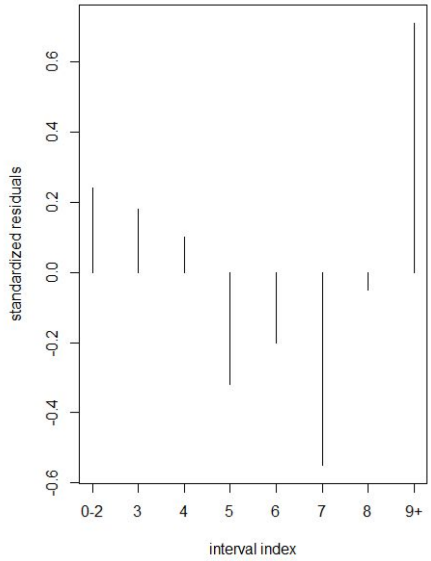
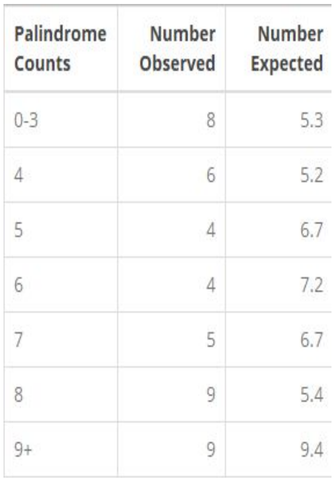


Figure 4.3.4. Standardized residual plot with an interval size of 4000.

Looking at the standardardized residuals plot in Figure 4.3.4, there are no residual values equal to ≥ 3, which indicates a good fit of the distribution.

Similarly, to look at the distribution from multiple perspectives, we solve the same methods with an interval size of 5000. This time $\frac{229,354}{5000} \simeq 46\ intervals$, which equates the lambda to be: $\lambda = \frac{296}{46} = 6.43$.



Palindrome Counts	Number Observed	Number Expected
0-3	8	5.3
4	6	5.2
5	4	6.7
6	4	7.2
7	5	6.7
8	9	5.4
9+	9	9.4

Table 4.3.5. Grouped distribution and expected values with interval size 5000.

The chi-squared test statistic equates to 6.86, which is much smaller than the calculated chi squared cutoff value with 7 - 2 degrees of freedom of 11.07. In addition, the p-value is 0.231, which is a low p-value but still greater than the significance value of 0.05.

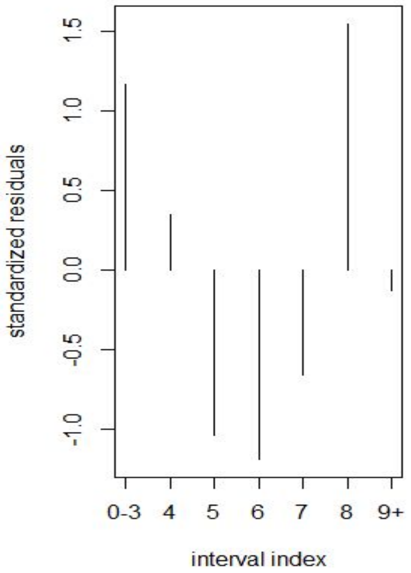


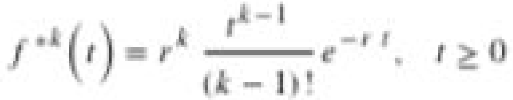
Figure 4.3.6. Standardized residual plot with an interval size of 5000.

Figure 4.3.6 shows the standardized residual values being between -1.0 and +1.5. Although higher than the previous interval size's residual values, the standardized residual values here do not exceed 3, which still allows us to still accept our null hypothesis.

d. The Biggest Cluster

It is imperative to take our findings back to the biologists so that they may arrive at a data-driven conclusion on how to handle the DNA. As part of the efforts to do so, we will formally answer the following question: Does the interval with the greatest number of palindromes indicate a potential origin of replication? We will also examine where the biggest cluster(s) lie. The intention of this section is to provide enough context to biologists about to start experimentally searching for the origin of replication.

It has been demonstrated that counts of palindromes segmented into intervals is approximated by the Poisson distribution. This fact gives direction on how to handle the data with the intention of isolating segments with the highest count of palindromes. An appropriate method to do so would be using the probability density function (pdf) of the Poisson distribution to compute the probability of a large cluster of palindromes, and from there we can examine if the interval with the largest number of palindromes can be described as a cluster (potential origin of replication). The Poisson process has the gamma pdf with shape parameter k and rate parameter r:

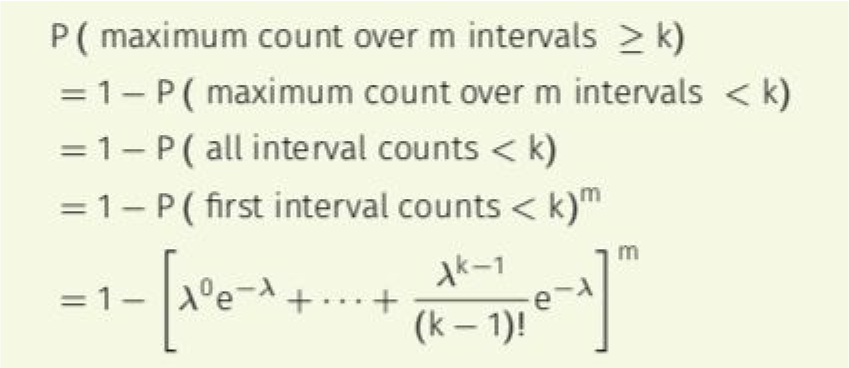


$$f^{*k}\left(t\right) = r^k \frac{t^{k-1}}{(k-1)!} e^{-r\,t}, \quad t \geq 0$$

Figure 4.4.1. Poisson probability density function.

The number of palindromes per interval of equal length are classified as independent observations. Given independent random variables, the location with the greatest number of palindromes per interval is the maximum. The null hypothesis here is that there are no unusually large clusters equivalent to the data following the Poisson Process. If the biologists empirically come to conclude the null hypothesis, there are no large clusters. If the alternative hypothesis is discovered, then there does exist large clusters. It is possible to compute the size of a cluster one can expect to see given that the data follows the Poisson process.

When considering how many palindromes could possibly fit in a single cluster, it is known that each interval has Poisson number of palindromes, and we have independent and identically distributed (iid) data across intervals, which makes finding the maximum easier. The data across intervals is iid under the assumption that within the control group palindromes occur with equal probability between regions. This allows the control group to be utilized to describe large clusters because if a number is observed to be larger than the maximum iid Poisson, it signifies to the biologists that they are in the presence of a large cluster. To mathematically represent this logic:



$$\begin{aligned}
&\mathrm{P}(\text{ maximum count over m intervals } \geq \mathrm{k}) \\
&= 1 - \mathrm{P}(\text{ maximum count over m intervals } < \mathrm{k}) \\
&= 1 - \mathrm{P}(\text{ all interval counts} < \mathrm{k}) \\
&= 1 - \mathrm{P}(\text{ first interval counts} < \mathrm{k})^{\mathrm{m}} \\
&= 1 - \left[\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}\right]^{m}
\end{aligned}$$

Figure 4.4.2. Derivation of the Poisson process.

From the above representation of the Poisson process model, m represents the number of intervals. This above expression represents the approximate probability that for a given estimation of lambda the greatest number of hits comes out to at least k. If this probability is small, a cluster of palindromes is probably larger than what is expected from the Poisson process. The maximum count of palindromes represents the test statistic in this circumstance (from right tail of the distribution), and the computation provides the p-value of the test statistic.

After programmatically running the described test statistic on different intervals, the majority of p-values are small (<0.05). This finding leads us to believe that there does exist a cluster of palindromes, which could be the origin of replications relevant to the biologists. We ran tests with the common assumption that alpha = 0.05. The intervals selected are 30, 60, and 90 and each interval has its own characteristic lambda and probability value. Since the model follows the Poisson distribution, as described above, lambda's value is estimated with MLE. Interval length is calculated by dividing the interval amount by total length. Maximum counts are generated by looking at max number of palindromes along all intervals, and interval displays the specific place with maximum count.

e. Alternative Hypothesis

As stated in the Background, CMV is a part of the herpes family and can be contracted throughout life. In the research paper "Cytomegalovirus Infection in Patients with AIDS", W.Lawrence Drew states that approximately 50% of the population will be seropositive for CMV by the age of 50 and are then carriers for life. This is not a deadly illness in adults unless combined with immunodeficiency, for example, HIV virus in which case can cause some serious diseases. Therefore, a deviance from normal HCVM igG antibody level could be an indicator of HIV infection. Using the dataset collected from rural Ugandan by Lisa Stockdale, our alternate hypothesis explores the difference of HCVM antibody between the HIV positive and negative population.

Taking the means of each CMV conditional on whether the person is HIV positive or not resulted in the following:

Mean of positive = 1.545

Mean of negative = 1.0173

Here we can see that HIV positive participants have a higher CMV antibody level than the HIV negative group. To further investigate this possible relation, a univariate logistic regression was run with 70:30 train-test ratio. This results in a p-value of less than 3e-09, and at a significance level of .05, we can assume that the results from Stockdale's test show statistical significance regarding HIV status having a possible negative impact on CMV antibody level. We can see through our prediction model that as CMV antibody level increases, so does the probability of being HIV positive. At varying values of CMV levels, the table below displays the big increase in chance.

CMV Level	Probability Of HIV
1	.0274
2	.240
3	.772

We can see that the chances of being HIV positive are fairly low when CMV antibody levels are low and when CMV levels are increased to 3, the chances of being HIV positive jump to 76%. It appears as though CMV levels could be an indicator for whether someone is HIV positive or not.

5. Theory

a. Homogeneous Poisson Process

In probability theory, the Homogeneous Poisson Process models randomly occurring phenomena (arrival times, decay times, relative position as part of a whole). The "process" of the model arises from the notion of points on a line distributed randomly, lacking signs of regularity in their placement.

Characteristic features of the process are as follows:

- The underlying rate λ where points (hits) occur and exists such that id doesn't change with location (homogeneity).
- The number of points falling in separate regions are independent.
- No two points land in exactly the same place.

The properties displayed above are enough to derive the formal model.

The poisson process serves as an adequate reference for making comparisons as it is a natural model for uniform random scatter. In our experiment, the homogeneous poisson process is the control model - meaning it represents palindromes occurring at random, without clusters. The strand of DNA can be conceptualized as a line, and location of palindrome as a point on that line. Uniform random scatter is saying that palindromes are scattered randomly and uniformly across the DNA. The number of palindromes in any small piece of DNA is independent of the number of palindromes in another non-overlapping piece. The chance that one small piece of DNA contains a palindrome is equivalent for all small pieces of DNA.

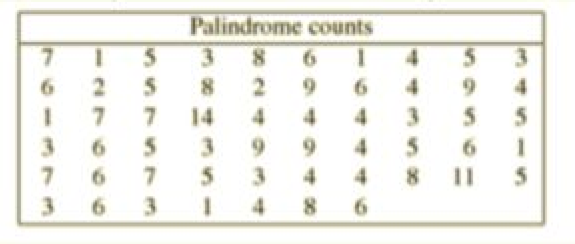
When exploring the question of how to simulate this random scattering, we recognize that this question is equivalent to simulating the control group. It is known that throws follow a uniform distribution on a given interval from the total possible interval length (0-296000). Then, you generate a simple random sample from this where chance is 1/0. Once you generate the control group, count for each interval how many observations fell into that interval. These observations should follow a Poisson distribution.

b. Chi-Squared Goodness of Fit Test

It is common to assume that observations are realizations of independent random variables (iid) from a specific distribution, such as the Poisson distribution. It is not assumed that data follows this distribution exactly, but that the distribution serves as a good proxy for randomness within the data. In the context of our study, if the Poisson distribution fits our data then it could be useful in locating unusual clusters based on finding small deviations in the data.

Homogeneous Poisson Process can be plausibly used as a reference model against which to seek an excess of palindromes. This methodology can only be sensibly followed if the model fits the data. If the model does not fit the data, this signifies too much heterogeneity in the locations of palindromes, so another model should be fitted.

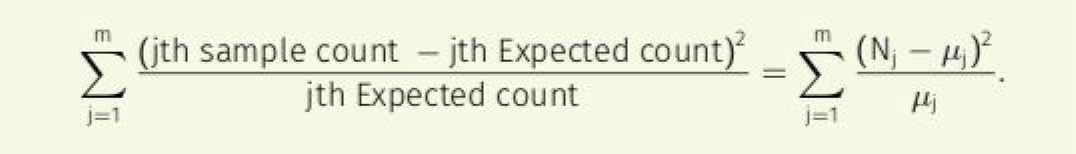
The next step in the process of determining whether our data can actually be said to fit the model is to apply the chi-squared goodness of fit test. DNA can be segmented into 57 non-overlapping intervals of length 4000. The number of complimentary palindromes in each segment should then be counted. The table below visualizes this process.



Palindrome counts									
7	1	5	3	8	6	1	4	5	3
6	2	5	8	2	9	6	4	9	4
1	7	7	14	4	4	4	3	5	5
3	6	5	3	9	9	4	5	6	1
7	6	7	5	3	4	4	8	11	5
3	6	3	1	4	8	6			

Table 5.2.1. Number of complimentary palindromes in a given DNA segment.

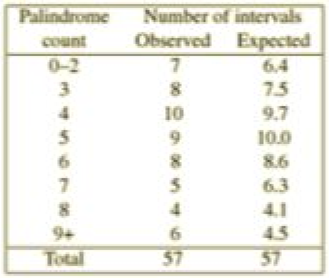
The number 4000 is arbitrarily chosen to make the number of observations recorded in the table reasonable. The last column of Table 5.2.3 contains expected number of segments continuing the specified number of palindromes as computed from the Poisson distribution. The categorized version is very useful as it informs the required computation to complete the chi-square test statistic (pictured below).



$$\sum_{j=1}^{m} \frac{(\text{jth sample count } - \text{jth Expected count})^2}{\text{jth Expected count}} = \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}.$$

Figure 5.2.2. Chi-Square test statistic. Nj is the observed count, muj is estimated by the Poisson (null) distribution with the estimated parameter (lambda hat).

From the total, it can be gathered that 57P (0, 1, or 2 palindromes in an interval of length 4000) = 57(e^-λ)[1+λ+(λ^2)/2]. The rate (actual value of λ) is unknown. Regardless, we can deduce a sample rate by understanding that there are 294 palindromes in 57 intervals of length 4000, which yields a sample rate of 5.16 per 4000 base pairs. By plugging in our estimated λ, we obtain 0.112 for the change in an interval of 4000 base pairs. The approximate expected number is 57 x 0.112 = 6.4.



Palindrome	Number of intervals	
count	Observed	Expected
0–2	7	6.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.3
8	4	4.1
9+	6	4.5
Total	57	57

Table 5.2.3. Displays the organized summary of palindromes and corresponding intervals.

Before applying the general form of the chi-square test statistic to the data, we will further elaborate on theory behind the test statistic and its implementation with the objective of clarifying its general applications, and specific purpose in our proposed setting of identifying palindromes in DNA sequences. Generally speaking, constructing a hypothesis test for a discrete distribution involves creating a distribution table representing the number of categories (or values) for the response and the number of observations that appear in each category. The generated counts are then compared to what would be expected under the null hypothesis, which in our case would be under the assumption that the data follows a Poisson distribution. When it comes to computing the probabilities, sometimes a parameter of the distribution needs to be estimated. We use the data to estimate these unknown values, measuring discrepancy between the sample count and the expected count. A large discrepancy upon computing the test statistic indicates that the data does not fit the distribution.

Given that in our data there exist 57 intervals and 294 palindromes, lambda hat is estimated by the average counts using method of moments and comes out to be 5.16. This informs the expected interval values reported in the table above. Applying our data to this understanding of the chi square test statistic, the computation flows as follows, and functions to compare observed data to expected:

$$\frac{(7-6.4)^2}{6.4}+\frac{(8-7.5)^2}{7.5}+\frac{(10-9.7)^2}{9.7}+\frac{(9-10)^2}{10}$$
$$+\frac{(8-8.6)^2}{8.6}+\frac{(5-6.3)^2}{6.3}+\frac{(4-4.1)^2}{4.1}+\frac{(6-4.5)^2}{4.5}=1.0$$

Figure 5.2.4. Implementation of the chi-squared test statistic for the given data.

Upon completing this computation, the next step is to determine the p-value, or significance level. P-values are always computed with chi squared distribution. Under the null hypothesis, the test statistic has an approximate chi square distribution with m-k-1 degrees of freedom, where m is representative of the number of categories and k is the number of parameters that are estimated to obtain the expected number of counts. Chi square distribution is a continuous distribution on the positive real line with a long right-tail density. As the degrees of freedom increase, the distribution starts to look symmetric and closer to that of a normal distribution.

$$P\left(\chi_6^2 \text{ random variable } \geq 1.0\right) = 0.98.$$

Figure 5.2.5 Computation of p-value for the given test statistic.

From the computation depicted above, it can be seen that the p-value is very large, so we fail to reject the null hypothesis. Therefore, we conclude that the data follows a Poisson distribution. The process that was concluded above represents the chi-squared goodness of fit test.

If the p-value is computed and its value is small, the fit is off. Plotting standardized residuals can help determine where the lack of fit occurs. To quantify this, follow the equation below for each category:

$$\frac{\text{sample count} - \text{Expected count}}{\sqrt{\text{Expectedcount}}} = \frac{N_j - \mu_j}{\sqrt{\mu_j}}.$$

Figure 5.2.6. Theory behind standardized residuals.

The denominator is transforming residuals so that they have approximately equal variance, enabling meaningful comparisons across categories as is required to meet the purpose of this computation.

Values of standardized residual larger than 3 indicate lack of fit. From our data, residuals of categories 7 and 9+ are large which indicates that this is where lack of fit occurs:

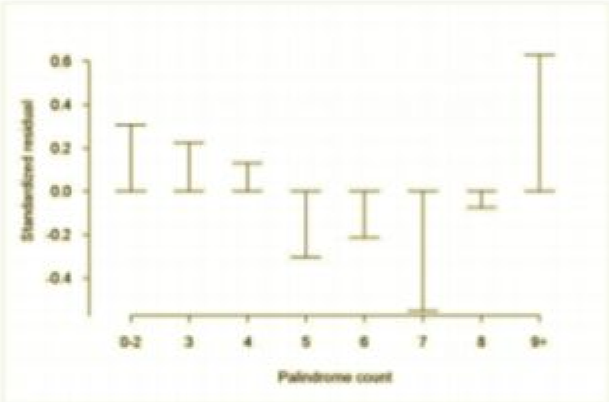


Figure 5.2.7. Plot of standardized residuals for palindrome counts over a standardized interval.

c. Choosing Interval Length

Choosing the right interval length is important as the correct interval length will allow us to make most of the data. With such a short interval length, the distribution of counts will contain many zeros and be highly skewed. For example we have an interval size of 500, which leads to 459 intervals, the total hits in from all intervals is low and the number of hits is also low, either 1s or 0s, making the count distribution not helpful.

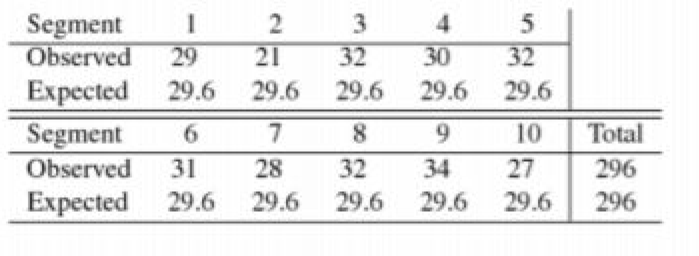
Similarly, on the other hand, with an interval size of 20,000, which equates to 11 intervals, each interval will hold lots of hits. This is a missplit in interval, as the 11 intervals will all have different palindrome counts, making it hard to analyze the visualizations or deduce anything from the distribution of counts.

Thus, in addition to choosing interval lengths between 4,000 and 5,000, we apply tests in multiple interval lengths within those ranges to double check.

d. Locations and Uniform Distribution

Under the Poisson process model, if the total number of hits in an interval is known, then the positions of the hits are uniformly scattered across the intervals. Thus, this allows us: first, to use the poisson process on a region to generate a random number, which represents the number of hits. And, second, generate locations for the hits according to the uniform distribution. So, under the uniform scatter distribution, the position of the palindromes are like 296 independent observations from a uniform distribution. Then, we can compare the 296 locations to the newly generated expected locations from the uniform distribution.

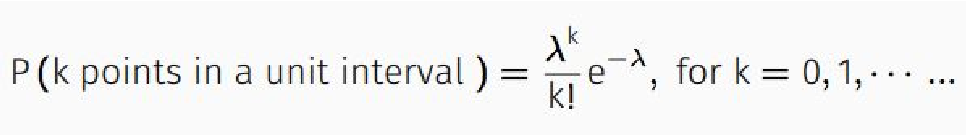
If DNA is split into 10 equal subintervals, then according to the uniform distribution each interval should contain 1/10 of the palindromes, as each 'observation' is independent from the other. With a chi square test, we can compare:



Segment	1	2	3	4	5	6	7	8	9	10	Total
Observed	29	21	32	30	32	31	28	32	34	27	296
Expected	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	296

e. Counts and Poisson Distribution

The characteristic features of the Poisson process are as follows: the count of points, called hits, in each region are independent of each other, the rate at which hits occur is the same at every location, and no two hits are in the same place. The Poisson Distribution counts the number of hits in each unit interval as follows:

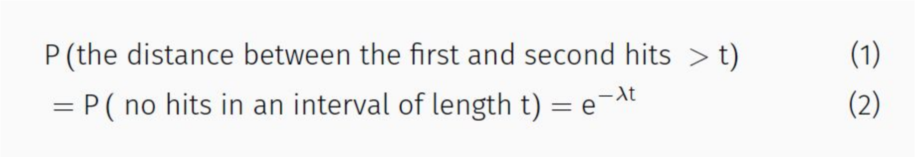


$$P(\text{k points in a unit interval}) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ for } k = 0, 1, \cdots \ldots$$

The expected value of the distribution is lambda, which is also the rate of hits per unit interval. The Poisson Distribution models the count of palindromes in equally divided intervals of the DNA.

f. Spacing and Exponential and Gamma Distribution

Distances between successive hits follows an exponential distribution with parameter lambda. That is,

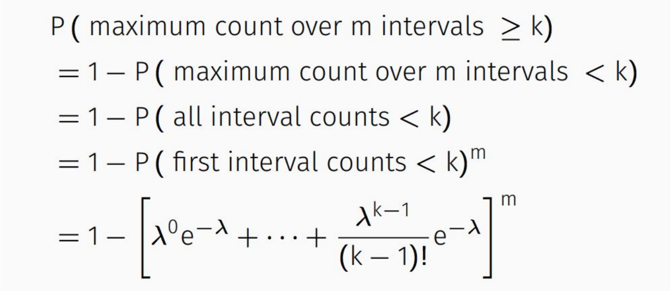


$$\begin{aligned} &P(\text{the distance between the first and second hits } > t) && (1) \\ &= P(\text{ no hits in an interval of length t}) = e^{-\lambda t} && (2) \end{aligned}$$

Additionally, distances between hits that are two palindromes apart from each other follows a gamma distribution with parameters two and lambda. The Exponential and Gamma Distributions can be used to model the number of base pairs between successive palindromes or palindromes that are two apart.

g. Maximum Number of Hits

The number of hits in a set of discrete intervals of equal length are independent observations from a Poisson distribution, under the Poisson process model. As a result, the greatest number of hits in a collection of intervals is the maximum of independent Poisson random variables. If the DNA is divided amongst *m* intervals, then we can derive the Poisson equation, shown in Figure 5.7.1.



$$\begin{aligned} &P(\text{ maximum count over m intervals } \geq k) \\ &= 1 - P(\text{ maximum count over m intervals } < k) \\ &= 1 - P(\text{ all interval counts} < k) \\ &= 1 - P(\text{ first interval counts} < k)^m \\ &= 1 - \left[\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}\right]^m \end{aligned}$$

Figure 5.7.1. Derivation of Poisson process model.

For a given estimate of λ, from the expression in Figure 5.7.1, the approximate chance that the greatest number of hits is at least k can be found. If this probability is abnormally small, then it indicates that there is a cluster that is larger than expected from the Poisson process. The maximum palindrome counts can be utilized as a test statistic, where Figure 5.7.1 computes the p-value for the test statistic, with a threshold of $\alpha = 0.05$.

h. Parameter Estimation

i. Method of Moments

Suppose we have an independent sample $X_1,...,X_n$ from a Poisson distribution with unknown rate parameter lambda (λ). The Method of Moments is an estimation technique that can be found the following way:

1. Find the expected value [E(X)] where X has a Poisson distribution with rate λ.
2. Express λ in terms of E(X).
3. Replace E(X) with $\bar{x}$ to produce an estimation of λ, called $\hat{\lambda}$.

For the Poisson distribution:

$$E(X) = \lambda \Longrightarrow \bar{X} = \hat{\lambda}.$$

Method of moments is based on the assumption that sample moments do provide good estimates of the corresponding population moments, therefore the average of the results obtained should be close to the expected value and only become closer as more trials are run. It is important to note that while the MME is consistent, it is not always efficient in achieving small mean square error.

ii. Maximum Likelihood

Suppose we have an independent sample $X_1,...,X_n$ from a Poisson distribution with unknown rate parameter lambda (λ). The Maximum Likelihood Method searches among all Poisson distributions to find the one that places the highest chance on the observed data. For the Poisson distribution, the chance of observing $X_1,...,X_n$ is:

$$\frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \times \cdots \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}e^{-\lambda} := L(\lambda)$$

For observed data, this is a function of λ that is also known as the likelihood function. Maximum likelihood estimates the unknown parameter by the λ-value that maximized the likelihood function.

Since the function is monotonically increasing, it is more useful to use the natural log of the likelihood function, denoted as l. This is allowed since ln(L) is maximized at the same value as L. In order to find the maximum, we look at the first-order derivative and set it equal to 0:

$$\frac{\partial}{\partial\lambda} l(\lambda) = \frac{\partial}{\partial\lambda}\left[\sum_i x_i \log(\lambda) - n\lambda - \sum_i \log(x_i!)\right] = \sum_i /\lambda - n = 0.$$

By solving the last equation for λ we obtain: $\hat{\lambda} = \bar{x}$.

Maximum likelihood relies on a large sample size for low bias. When the sample size is larger than 30, the MLE should be unbiased and normally distributed. If the sample size is small, the MLE can be highly biased.

i. Assessing Parameter Estimation

i. Mean Square Error:

To compare parameter estimates, mean squared error is defined as:

$$\mathrm{MSE}(\hat{\lambda}) = \mathbb{E}(\hat{\lambda} - \lambda)^2 = \mathrm{Var}(\hat{\lambda})\text{variance} + \left[\mathbb{E}(\hat{\lambda}) - \lambda\right]^2 \text{squared BIAS}$$

Many of the estimators are unbiased, however, sometimes estimator with a small bias will have a small MSE.

ii. Asymptotic Distribution

The theorem states that under certain conditions, as the sample size increases, the maximum-likelihood estimator satisfies:

$$\hat{\lambda} \rightarrow \lambda$$

$$\hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{1}{n\mathrm{I}(\lambda)}\right)$$

l(λ) is called the Fisher’s Information Matrix, which is defined below.

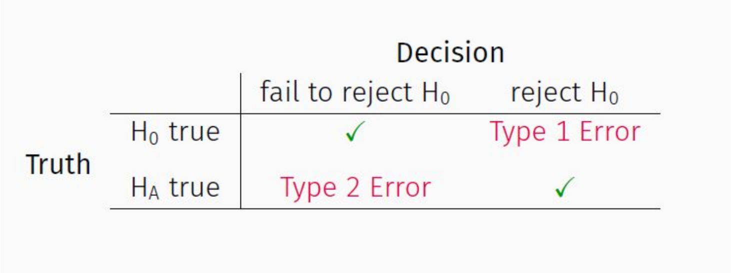
$$\mathrm{I}(\lambda) = \mathbb{E}\left(\frac{\partial}{\partial\lambda} \log f_\lambda(X)\right)^2 = -\mathbb{E}\left(\frac{\partial^2}{\partial\lambda^2} \log f_\lambda(X)\right).$$

Thus as n increases, the MLE estimator will have an asymptotically normal distribution and this distribution can be used to build confidence intervals for unknown λ (Both shown below).

$$\sqrt{n\mathrm{I}(\lambda)}\left(\hat{\lambda} - \lambda\right) \sim \mathcal{N}(0,1). \qquad \hat{\lambda} \pm 1.96\sqrt{n\mathrm{I}(\lambda)}.$$

j. Hypothesis Test

The chi-squared goodness of fit test and the test for the maximum number of palindromes in an interval are both examples of hypothesis tests. This section will explore a hypothesis test that uses parameter values. When using hypothesis tests the researchers must assume the null hypothesis is true and analyze data to determine how likely it is to be true. The outcome of researchers decisions can often be affected by the introduction of Type I and Type II errors. Before choosing to accept or reject the null, researchers must take into account the values of Type I and Type II errors. Type I errors, denoted as alpha, alter the significance of the test. They lead researchers to falsely reject the null when it is actually true. Type II errors, denoted as beta, alter the power of a test. They cause researchers to falsely fail to reject the null when the alternate hypothesis is true. This can be seen in the chart below:

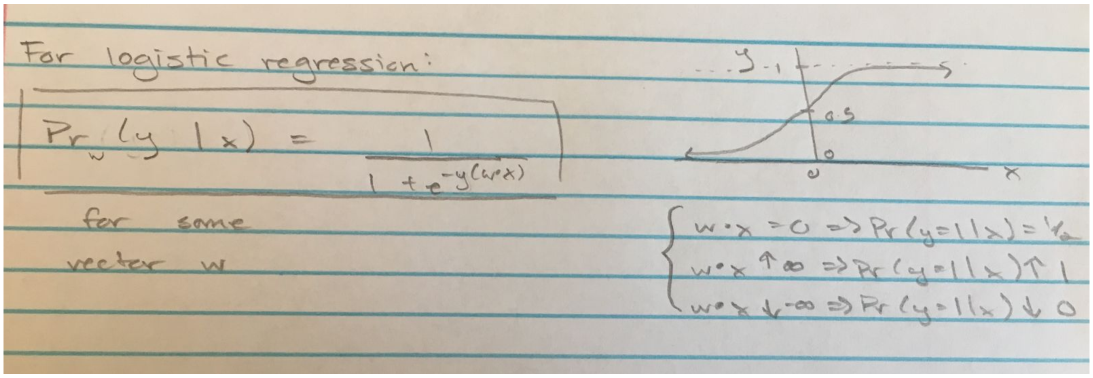


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

Typically, the value of the Type I error is set in advance, while the Type II error needs to be computed for various values. The power of a test is determined by 1-beta and the higher the power of a test, the better the test is.

k. Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, independent variables. There must be two or more independent variables in order to use this method. Simple logistic regression is analogous to linear regression, except that the dependent variable is nominal, not a measurement. Because the dependent variable is binary, the probability of y = 1 is demonstrated in Figure 5.11.1. As the dot product of a vector *w* and *x* approach infinity, *y* approaches 1 whereas if *w* and *x* approach negative infinity, *y* approaches 0. The vector *w* is calculated so that the logistic function fits the data well. This will maximize the probability function and minimize the loss function, given the training dataset.



For logistic regression:

$$\boxed{Pr_w(y \mid x) = \frac{1}{1+e^{-y(w \cdot x)}}}$$

for some vector w

$$\begin{cases} w \cdot x = 0 \Rightarrow Pr(y=1 \mid x) = 1/2 \\ w \cdot x \uparrow \infty \Rightarrow Pr(y=1 \mid x) \uparrow 1 \\ w \cdot x \downarrow -\infty \Rightarrow Pr(y=1 \mid x) \downarrow 0 \end{cases}$$

Figure 5.11.1. Diagram of logistic regression.

6. Conclusion

Throughout this study, we showcased the distributions of various aspects of the data sequence of CMV by dividing the 229354 bases into 42 intervals to observe the counts of palindromes. We concluded that the palindrome data follows a Poisson distribution. In addition, the spacings between consecutive palindromes and the spacings between consecutive pairs of palindromes follow exponential distribution and the spacings of consecutive triplets of palindrome locations follow a gamma distribution. The distribution of palindrome counts can be approximated with a Poisson distribution. Within the data, we found that there is a cluster of palindromes, which could be the origin of replications.

Works Cited

1. W. Lawrence Drew. “Cytomegalovirus Infection in Patients with AIDS.” *Clinical Infectious Diseases*, vol. 14, no. 2, 1992, pp. 608–615. *JSTOR*, www.jstor.org/stable/4456333
2. Jelena Bradic, MATH 189 (Data Analysis & Inference) Slides, Winter 2019