Identifying Palindromic Patterns in DNA

Math 189/289C

Homework 3 Report

Feb. 23rd, 2017

By:

- Emma Roth, 4th Year Computer Science/Bioinformatics Major
    - Introduction, Chi Square Theory, Conclusion
- Ileena Mitra, 1st year Bioinformatics PhD
    - Section 4, Conclusion
- Megan Lee, 3rd Year Mathematics/Economics Major
    - Introduction, Section 2
- Jing Gu, 2nd year Chemistry MS
    - Theory, Section1
- Keven Nguyen, 4th Year Mathematics/Economics Major
    - Theory, Section 3
- Adithya Bharadwaj Balaji, 1st year Graduate student in Electrical and Computer Engineering
    - Hypothesis, Theory, Section 3, Section 4

INTRODUCTION

Cytomegalovirus (CMV) is a congenital viral infection. It is currently the most prevalent congenital viral infection amongst developed countries: 1/3 of all children have been infected by age 5, and over half of all adults have been infected by the age of 40 (CDC). This virus is transmitted through body fluids, and most people never show symptoms. Those who do present symptoms usually present fever, sore throat, fatigue, and/or swollen glands. Once a person is infected with CMV, it remains in their body for the rest of their life and can reactivate at any time. Though it is quite common, cytomegalovirus generally is not symptomatically manifested in healthy people infected; however, in patients with already compromised immune systems, it can be life threatening. For infants in particular cytomegalovirus can lead to "sensorineural hearing loss (SNHL), mental retardation, microcephaly, developmental delay, seizure disorders, and cerebral palsy" (Swanson 2013). Given its prevalence worldwide, the lack of widespread knowledge about cytomegalovirus makes it a relevant area of study as there is currently no cure for the infection.

The literature surrounding cytomegalovirus emphasizes the important relationship between the virus and the immune system. Present studies aim to understand the humoral and cellular response to the virus "in order to finally identify CMV-protective antigens and to acquire in-depth understanding of the immune mechanisms" of the virus (LaRosa 2012). Currently, CMV is treated via antiviral drugs. These drugs aim to prevent CMV infection and symptoms, and have been shown to be effective particularly in organ and stem cell recipient patients. This antiviral therapy has vastly improved treatment management for patients with compromised immune systems (Ahmed 2011). However, as with any drug, limitations in the efficacy of treatment of cytomegalovirus persist. To further understand this virus, scientists have sought to study the way in which it replicates, which has in turn led to the study of

palindromes.

A palindrome refers to a sequence of nucleotides in a strand of DNA or RNA that contains its own converse complements. Two viruses from the same family as CMV, Herpes simplex and Epstien-Barr virus, have origins of replication identified by palindromic sequences of nucleotides. The origin of Herpes simplex is marked by a palindrome of 144 nucleotides, while the origin of Epstien-Barr is marked by a cluster of short, palindromic sequences. In 1992, professors at Stanford analyzed the frequency and complexity of repeats across the human CMV genome sequence in the hopes of using patterns to identify its origin of replication. Using statistical methods on the genome, they found a particular region of interest (between locations 92,100 and 93,500) in the genome consisting of the greatest frequency and density of repeats (Masse et al). Biological experiments proved this region was, in fact, the origin of replication by cutting this section out of the genome and observing if it could replicate on its own. This study is an example of the important role that identifying patterns in genomes plays gaining further insight into the cytomegalovirus. In this report, we use data on the locations of palindromes in the CMV genome to identify clusters of patterns that are biologically significant. We find that the region that is the most likely candidate to contain the origin of replication is between locations 90,000-95,000 in the CMV genome.

HYPOTHESIS

We aim to identify the distribution of a cluster of palindromes and whether they are randomly distributed or follow a certain distribution, for example, a normal distribution, poisson distribution or uniform distribution. If the distribution of the clusters indicates that they do not occur by a random chance and follow a certain distribution, then these clusters might signify the location of the origin of the replication. This can help scientists begin investigation at these locations, thereby saving time and money in the search for replication of the Human Cytomegalovirus (CMV).

In this case study, we focus on the unusual clusters of complementary palindromes. We are given the location of the occurrences of complementary palindromes of length between 10-18 in the CMV DNA sequence. The CMV DNA is 229,354 base pairs long and since this consider only palindromes of length 10-18, this comes out to be 296 number of palindromes.

We propose a null hypothesis H_0,

H_0 = The palindromes are randomly scattered (follow a random distribution)

H_A = The palindromes are not randomly scattered (follow a certain distribution like poisson, normal, uniform, or gamma etc.)

DATA

The data we analyzed is a collection of the starting locations of palindromes in the DNA sequence of CMV. These locations were found by pattern search algorithms created and implemented by Leung et al. in 1991 on the CMV DNA sequence discovered by Chee et al. in 1990 (Leung, Chee). Any palindrome that consisted of less than 10 letters was excluded. In total, the CMV DNA is 229,354 letters long, and our dataset consists of the 296 locations of the beginnings of palindromic sequence. We will use these locations to attempt to find clusters of palindromes within the DNA sequence and attempt to find a cluster that represents the origin of replication in CMV.

THEORY

1. The Homogeneous Poisson Process, Applied probability modeling

The Homogeneous Poisson Process

The Poisson process is a model that represents discrete arrivals such as arrival times of telephone calls and number of palindromes in the DNA sequence. N(t) notes for the number of arrivals that have occurred from 0 to T. The Homogeneous Poisson process has the rate of arrival per unit time to be the same across the entire time periods. There are some characteristics for the Homogenous Poisson process.

1) Rate λ does not change with location or time.

2) In an infinite small time interval (dt), there can only be 1 or 0 arrival.

3) The number of arrivals in one interval is independent from number of arrivals that fall into other intervals.

4) The time between two consecutive arrivals are independent and follows exponential

distribution.

N(t) follows poisson distribution with $\mu = \lambda t$, where μ is the expected number of hits per unit interval.

$$P(N(t) = k) = \frac{\mu^k}{k!} e^{-\mu} \text{ for } k = 0, 1, 2, \ldots$$

Usually λ is unknown, which needs to be estimated. There are two common methods to find estimators, maximum likelihood estimator (MLE) and method of moments.

Poisson Process Simulation through

(a) Uniform distribution

- N(t) is simulated via the arrival time (Ti), the moments when i th event comes.

Step1: N = Poisson(λt)

Step2: Generate N random variables Ui ~ Unif(0,1)

Step3: (T1, T2, T3…Tn) = t(end time point) * sort(U1, U2, …Un)

(b) Exponential distribution

- The inter-arrival times $Wi = T_i - T_{i-1}$

Step1: $T_0 = 0$

Step2: For i = 1, 2, 3…n, do

E = Exp(λ);

$T_i = T_{i-1} + E_i$

2. Chi-Square Goodness-Of-Fit Test, Hypothesis Tests

Hypothesis Tests

The chi square goodness-of-fit test and test for the maximum number of palindromes in an interval are two examples of hypothesis tests. A hypothesis test is a statistical test that is used to determine if there is enough evidence in a sample to infer a certain condition is true for the population. The hypothesis test examines two opposing hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is the statement we are testing, and the alternative hypothesis is a

statement we would like to be proven true. As mentioned previously, our H_0 is that the palindromes are randomly scattered, and our H_A is that the follow a certain distribution.

Hypothesis tests are not designed to select the more likely of the two hypotheses and they must be mutually exclusive. A null hypothesis can only be rejected if statistics have proven there are enough data to support the alternative hypothesis.

Type I Error, Type II Error, and Power

There are several types of error in statistical hypothesis testing. Type I error is the incorrect rejection of a null hypothesis, and can be called a “false positive”. Type I error rate, also called significance level, is the probability of rejecting the null hypothesis given that it is true. It is represented by the Greek letter α and is set before the test, usually to 0.05 (5%), stating that it’s acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

Type II error is the incorrect acceptance of the null hypothesis, or a "false negative”. It is represented to Greek letter β, and related to power of a test. Typically, β is calculated for various values of the alternative hypothesis. The risk of committing a type II error is decreased by ensuring that your test has enough power. Power (also known as sensitivity) of a hypothesis test is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. It measures the ability of a test to detect an effect if the effect actually exists, and is calculated by 1-β. High power indicates a good test.

	Null Hypothesis	
Decision	True	False
Fail to reject	Correct Decision (probability = 1 - α)	**Type II Error** - fail to reject the null when it is false (probability = β)
Reject	**Type I Error** - rejecting the null when it is true (probability = α)	Correct Decision (probability = 1 - β)

Chi Square Goodness-of-Fit Test

We use a chi square test to determine if the observed data fits well with or varies significantly from a theoretical distribution. A chi square goodness of fit test is appropriate for categorical data with a random sampling method, and the sample observations at each level (bin) is at least 5. To compare observed data to expected, we

generate simulation data and compare it to the observed data. We then divide both the simulation data and the observed data into an equal number of bins of equal range, and use this to calculate the test statistic. The test statistic is calculated using the formula

$\chi^2 = \sum_{i=1}^{n} \frac{(Observed\ count - Expected\ count)^2}{Expected\ count}$, where n is the number of bins, observed count is the number of data points we observed in that bin, and expected count is the count of data points from the simulated data in a bin of the same range. The value obtained with this calculation is then compared to the established critical value obtained from the chi square distribution with n-1 degrees of freedom (n referring to number of bins) at a certain significance level. Degrees of freedom are broadly defined as the number of observations that are free to vary when estimating statistical parameters. The degree of freedom is equal to the number of bins -1 because if we know how many points are in the n-1 bins, then we know that the value in the nth bin is m- # of data points in n-1 bins. Significance level is discussed more below, but most often the significance level is set to either 0.1 (10%), 0.05 (5%), or 0.01 (1%).

To determine whether or not you can reject the null hypothesis, you compare your obtained χ^2 value to the χ^2 from the chi square distribution corresponding to the same number of degrees of freedom as your data set. This χ^2 value is associated with a p-value. You can reject the null hypothesis when the p-value associated with the χ^2 for your observed data is less than the set significance level

The Chi square goodness-of-fit test is used to compute the chance of observing a test statistic at least as large as ours under the random scatter model. A large test statistic indicates a lack of fit of the distribution.

P-Value

Assuming that the data are generated from the hypothesized distribution, we can compute the chance that the test statistics would be as large, or larger than that observed. This chance is called the observed significance level, or p-value. If the p-value is small, there is a reason to doubt the fit of the distribution. Computing the p-value requires a test statistic, or a scalar function of the observations, that summarizes the data. T-tests and Z-tests are quite common ways to examine if the data follows a normal distribution, however, the data we are studying does not follow a normal distribution. In this case, we can use chi square goodness-of-fit test to determine if our data fits a theoretical distribution (Poisson, Exponential) with a certain level of significance.

3. Uniform Distribution, Exponential and Gamma Distributions

For a homogeneous poisson distribution, the chance of occurrence of hits at the locations given by the exponential distribution, are uniformly distributed.

For example, suppose we are told that exactly one event of a Poisson process has taken place by time t, and we are asked to determine the distribution of the time at which the event occurred. Since a Poisson process possesses stationary and independent increments, it seems reasonable that each interval in [0,t] of equal length should have the same probability of containing the event. In other words, the time of the event should be uniformly distributed over [0, t].

The uniform distribution gives a constant probability and its probability density function is a continuous uniform distribution on the interval [a, b]. It is given by,

$$P(x) = \frac{1}{b-a}, \qquad \text{for } a \leq x \leq b$$

The spacing between the palindromes can be distributed either uniformly, exponentially or in a gamma distribution. In the case of a homogeneous poisson process, the distances between successive hits follows an exponential distribution. This is given by,

$$P(\textit{the distance between the first and second hits} > t)$$

$$= P(\textit{no hits in an interval of length } t) = \lambda e^{-\lambda t}$$

The characteristics of an exponential distribution are given as follows:

Probability distribution function = $\lambda e^{-\lambda t}$ for $t \geq 0$

Mean= $1/\lambda$

Variance= $1/\lambda^2$

where $\lambda > 0$ is the parameter of the exponential distribution.

The gamma distribution has two positive real numbered parameters that are

parametrized in the following three different manners:

1. With shape parameter k and scale parameter θ.

2. With shape parameter $\alpha = k$ and an inverse scale parameter $\beta = \frac{1}{\theta}$, called the rate parameter.

3. With shape parameter k and mean parameter $\mu = \frac{k}{\beta}$

When specified in terms of and , we have the density function

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

We can observe that the exponential distribution and the chi-squared distribution are special cases of gamma distribution with k=1 and θ=1, and k=N/2 and θ=2 σ^2 respectively. The gamma function is great for modelling random waiting times between occurrences.

4. Parameter Estimation and Properties of Parameter Estimates

Given the probability density function of the Poisson distribution:

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

We can estimate the parameter, λ, or the hit rate, by taking the empirical average number of hits per unit interval. We prove this by first showing that the expected value of a Poisson random variable is λ.

$$E(X) = \sum_{0}^{\infty} x \frac{e^{-\lambda}\lambda^X}{X!}$$

$$= \sum_{1}^{\infty} x \frac{e^{-\lambda}\lambda^X}{X!}$$ since the x = 0 term is itself 0

$$= \sum_{1}^{\infty} \frac{e^{-\lambda}\lambda^X}{(X-1)!}$$ divide top and bottom by x

$$\lambda e^{-\lambda} \sum_{1}^{\infty} \frac{\lambda^{X-1}}{(X-1)!}$$ factor out constants

$$= \lambda e^{-\lambda} \sum_{0}^{\infty} \frac{\lambda^X}{X!}$$

$$= \lambda e^{-\lambda} \, e^{\lambda}$$

$$E(X) = \lambda$$

Since the expected value of a Poisson random variable is indeed λ, the empirical average is a good estimate for λ.

RESULTS

1. Random scatter

Given 296 palindromes were found with length of at least 10 letters long, we used histogram to segment the DNA chain into intervals with length of 4000 base pairs and then count the number of palindromes found in each interval. Figure 1 suggests that there seems to be clusters of palindromes around locations 74,000, 94000, and 19,400 of CMV DNA. We further compared histograms of the actual palindromes with those based on randomly generated numbers to see if the outliers on the DNA are unusual. he histograms of two sets of random scatters do not show distinct high spikes as patterns for clusters (Figure 2 and 3). Therefore, we hypothesize that clusters at those locations are atypical, which are likely to be the locations for origin of replication.

Figure 1: Histogram for locations of palindromes in DNA

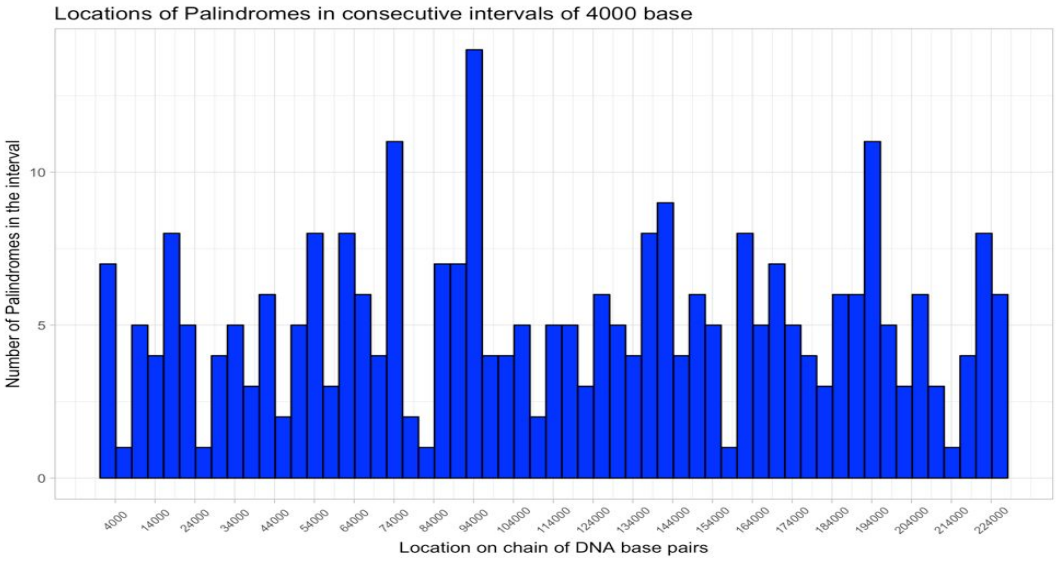


Figure 2: Location of hits in simulated random data set 1

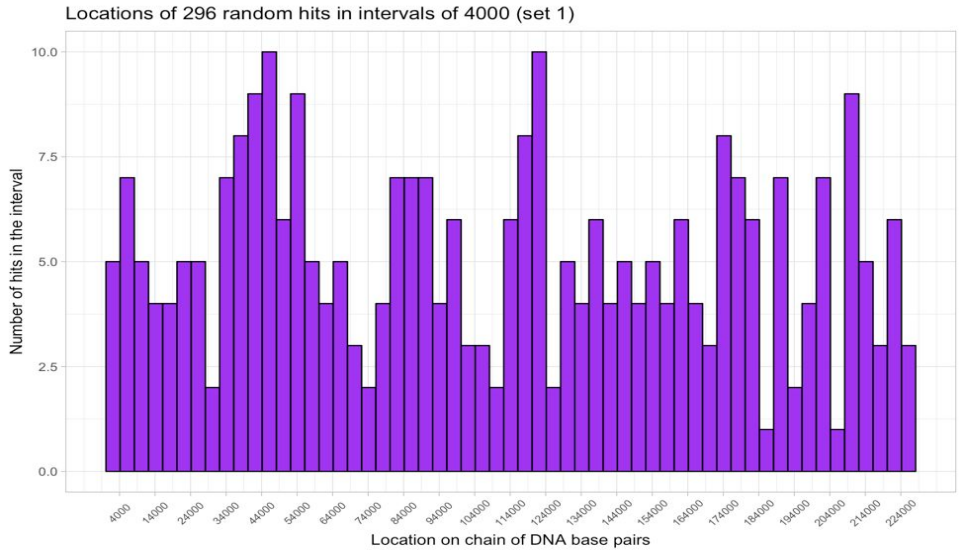
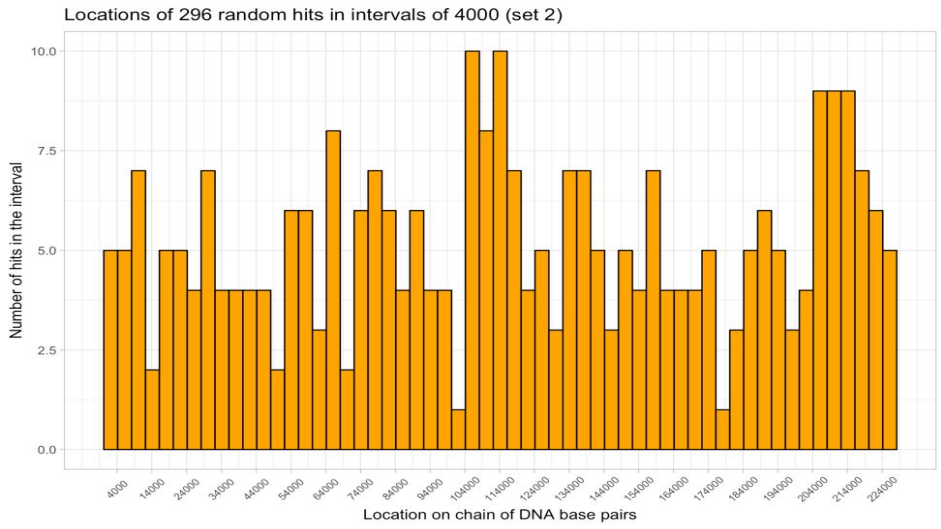
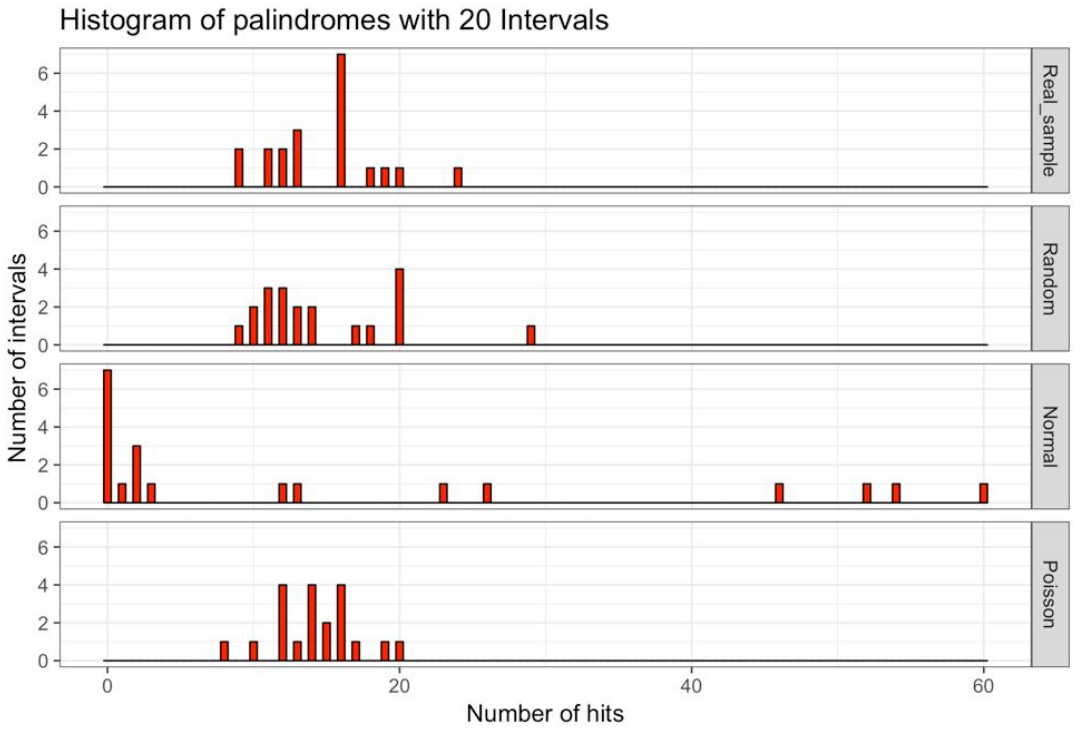


Figure 3:Location of hits in simulated random data set 2



We also examined data sets simulated from normal and Poisson distributions. The procedure to simulate of Poisson distribution is addressed in the above theory section. We then plotted histograms of number of palindromes in total of 20 intervals for each data set. When the length of intervals increases, one may observe bigger difference among the data sets since clusters tend to be more easily spotted when looking at bigger intervals. From Figure 4, we can see that real sample tends to have a closer distribution to data sets from Poisson distribution than the normal distribution. Though histograms for number of palindromes are helpful to visually spot occurrence of clusters, we also need to compare locations of the palindromes and examine the spacings between consecutive palindromes. More formal statistical tests in the following sections are required to determine whether a cluster is a potential replication site.

Figure 4: Comparison of number of palindromes/hits in 20 intervals for different distributions



2. Locations and spacings

In this section we seek to examine the spacing between consecutive palindromes, the sum of consecutive pairs spacing, and the sum of consecutive triplets spacing.

The first graphical method we attempt to use to detect significant findings in the spacing is a scatterplot.

Figure 5: Scatter plot of spacing between consecutive palindromes

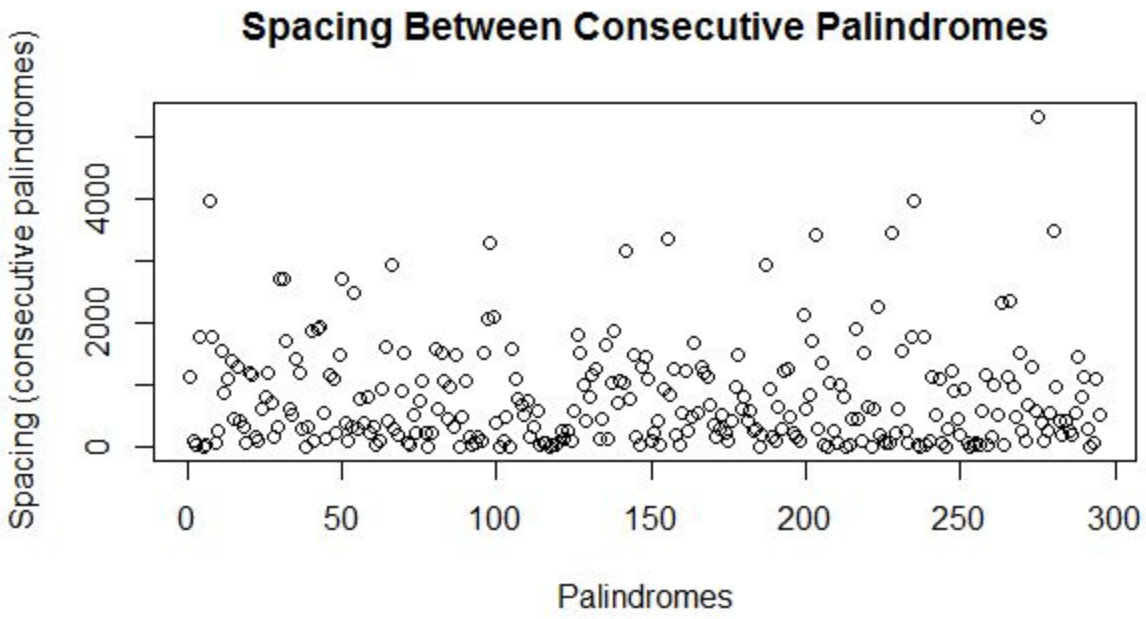
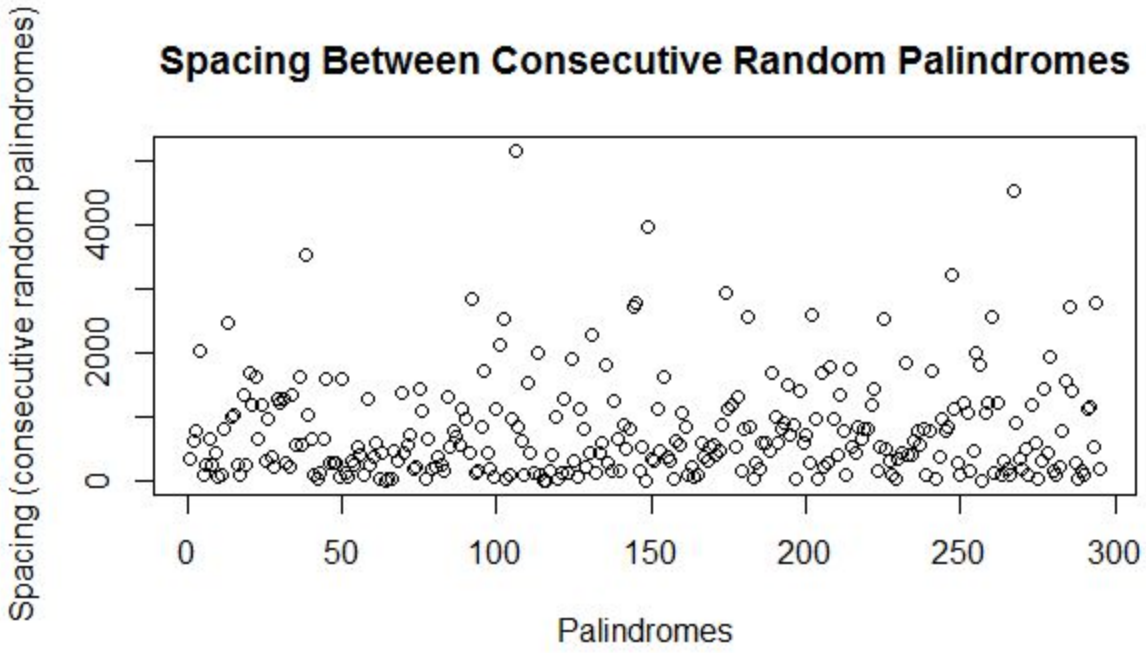


Figure 6: Scatter plot of consecutive spacing in random data



Upon comparing Figures 5 and 6, we can detect no patterns that could provide us with significant information. As a result we turn to other graphical measures.

A histogram is a useful and appropriate graphical method to accomplish this task since we have a large volume of data samples and we seek to identify unusual clusters of palindromes. The histograms presented provide us with significant information regarding the distribution of spacing between different groupings of palindromes. We also compare our findings from the sample data with the randomly generated data created in the previous section to see how accurately it models the sample data.

First we examine the spacing between consecutive palindromes. From the sample data and the randomly generated data, this spacing was found by subtracting successive palindromes from one another. For instance, given two consecutive palindromes and their starting locations X_1 and X_2, the spacing between the two consecutive palindromes is calculated by taking the difference X_2-X_1. This was done for all 296 samples from the two data sets.

Figure 7: Histogram of spacing between consecutive palindromes

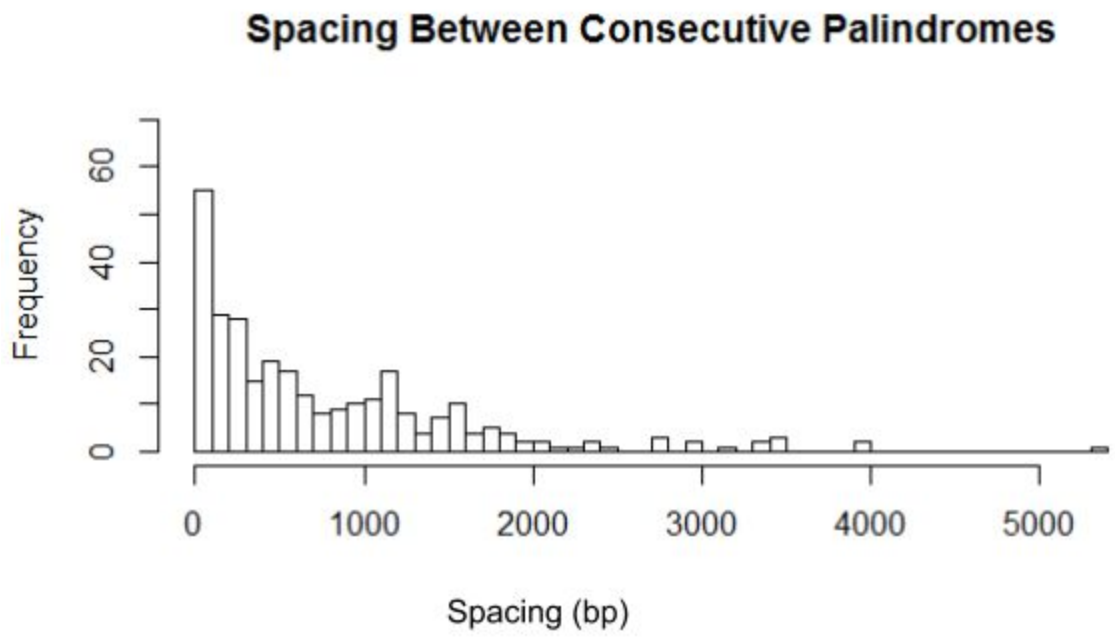
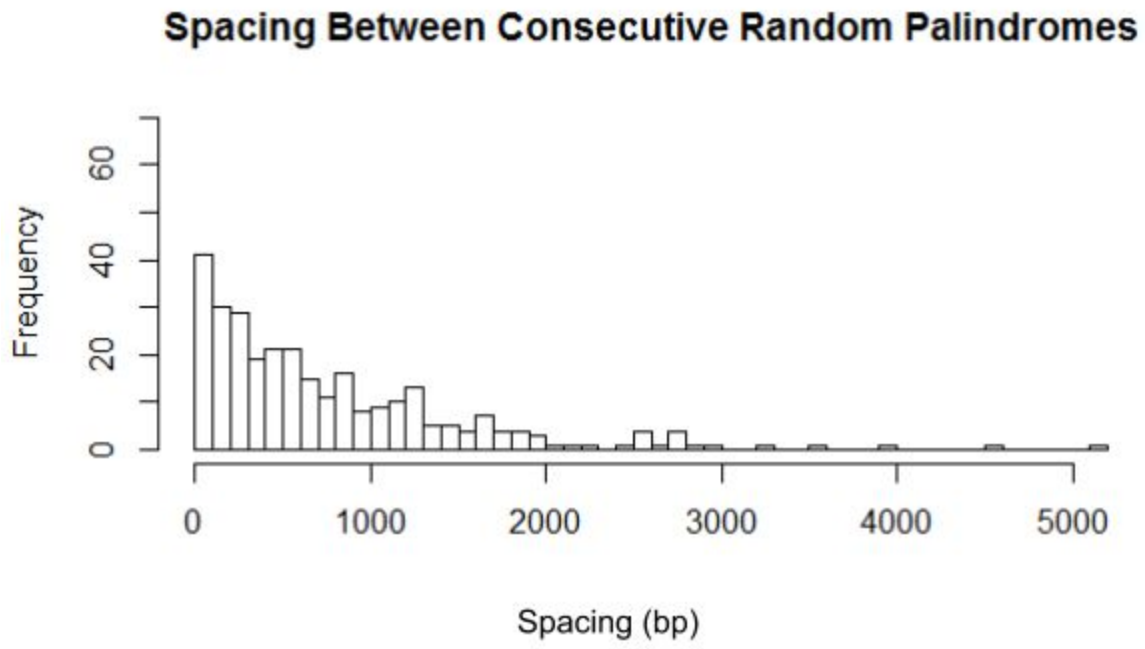


Figure 8: Histogram of spacing between random consecutive palindromes



Comparing figures 7 and 8 reveals that the randomly generated data closely follows the given data. A noticeable difference between the two sets of data is that there are less low values of spacing in the random data as compared to the given data. This indicates that there is more clustering in the real data than in the random data.

Repeating this procedure to examine the spacings of consecutive pairs and triplets between the given and randomly created data produces the following histograms:

Figure 9: Histogram of spacing between consecutive palindrome pairs

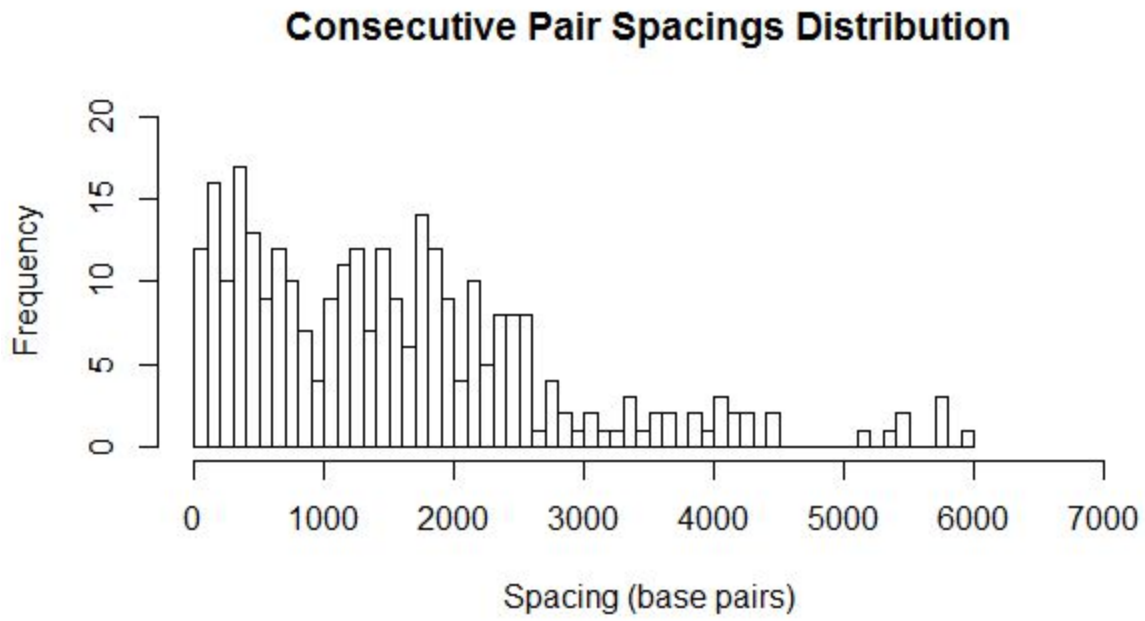


Figure 10: Histogram of spacing between consecutive random palindrome pairs

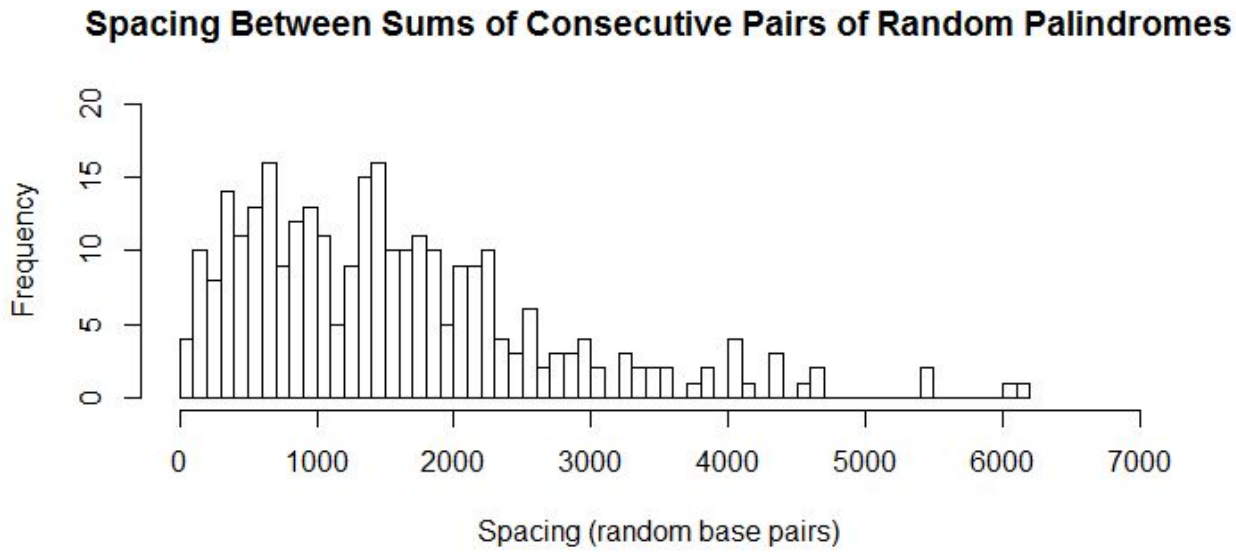


Figure 11: Histogram of spacing between consecutive palindrome triplets

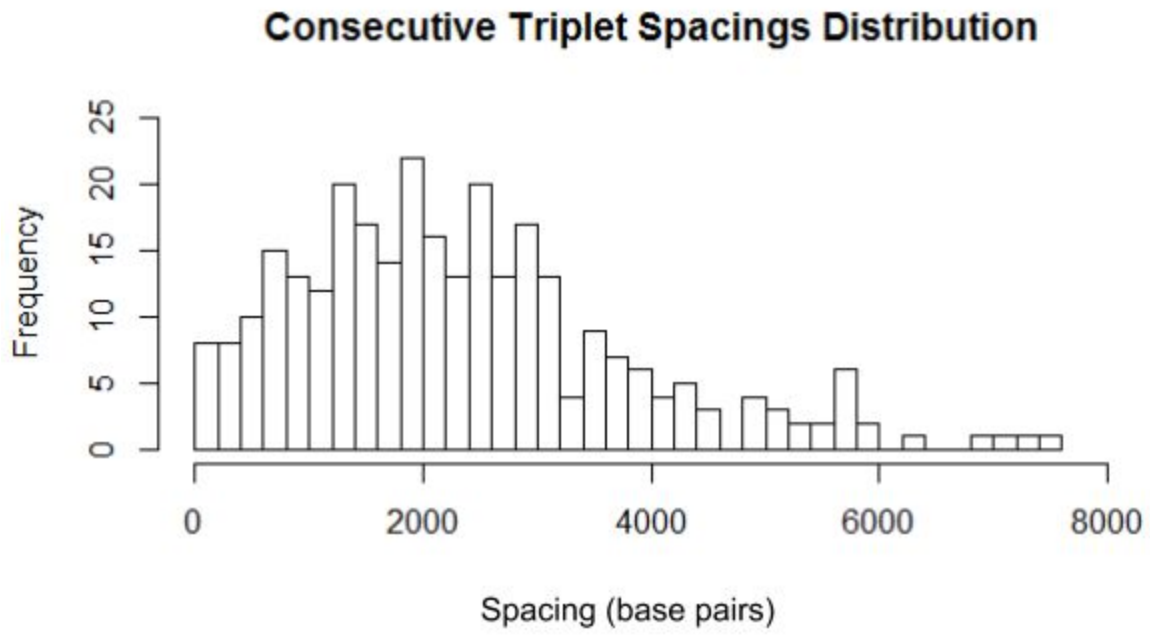
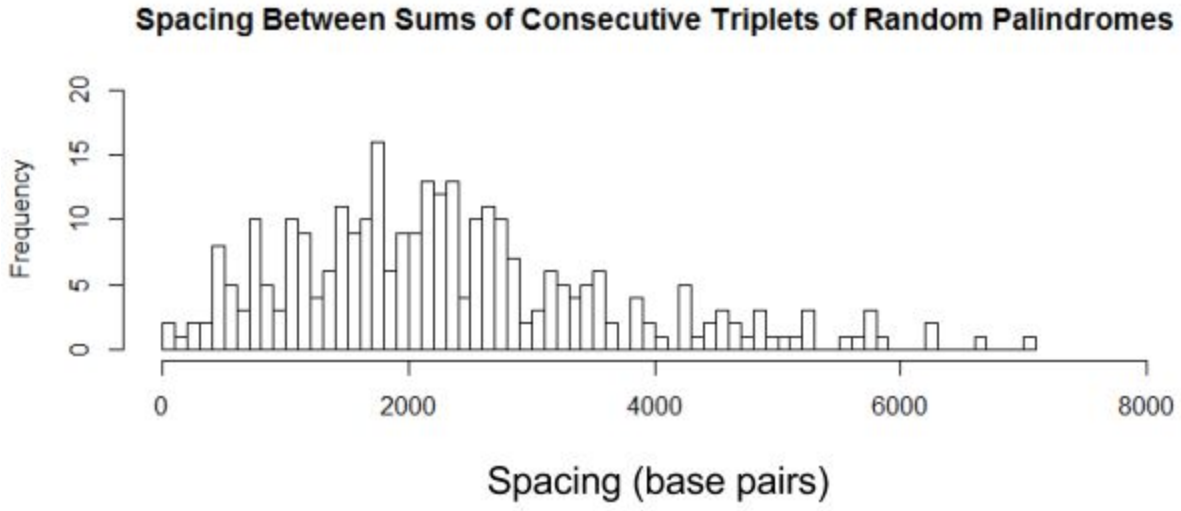


Figure 12: Histogram of spacing between random consecutive palindrome triplets



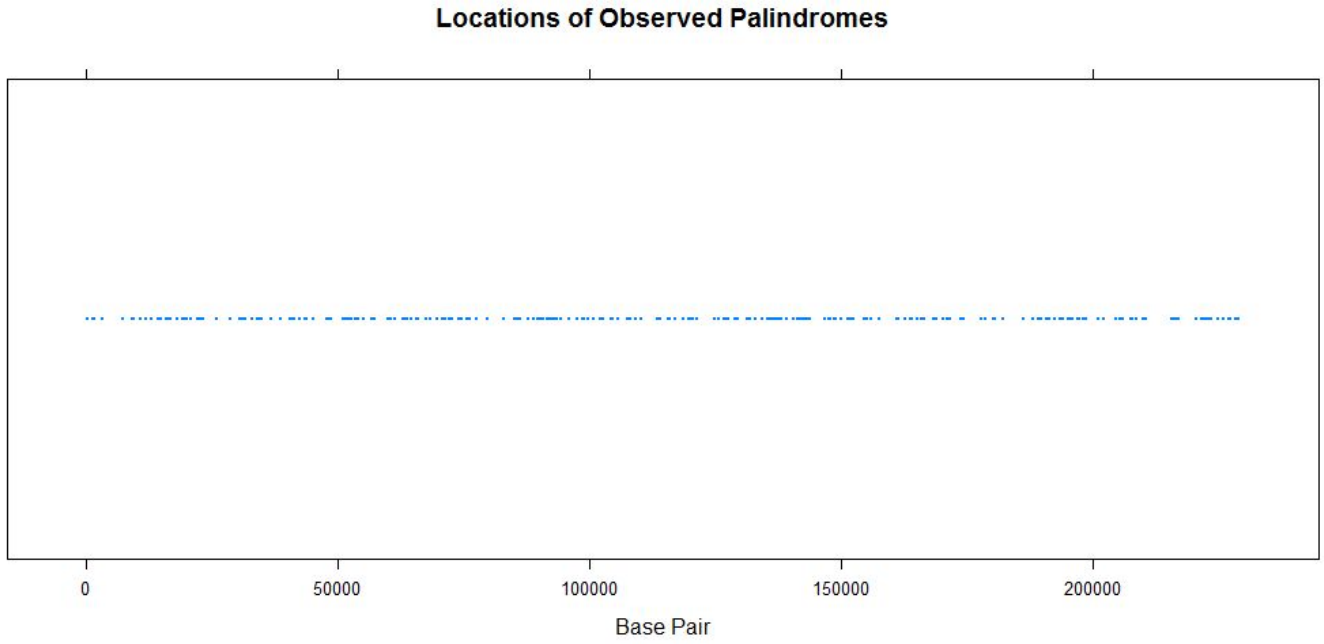
Comparing successively the spacings between consecutive palindromes, consecutive pairs of palindromes, and consecutive triplets of palindromes between the given and randomly created data shows that the randomly created data is generally a good model for the data, however the real data reveals more clustering than the

random data does. Where the two data sets repeatedly diverge is that in the random data there appears to be large clusters of palindromes with larger spacings in between them, whereas in the given data there are large clusters of palindromes with smaller spacings in between them. These findings suggest that the data is not completely random and that clusters of palindromes do exist. As we seek to identify clusters of patterns in the data through more formal statistical testing in the following sections we will obtain a better understanding of the data.

3. Counts

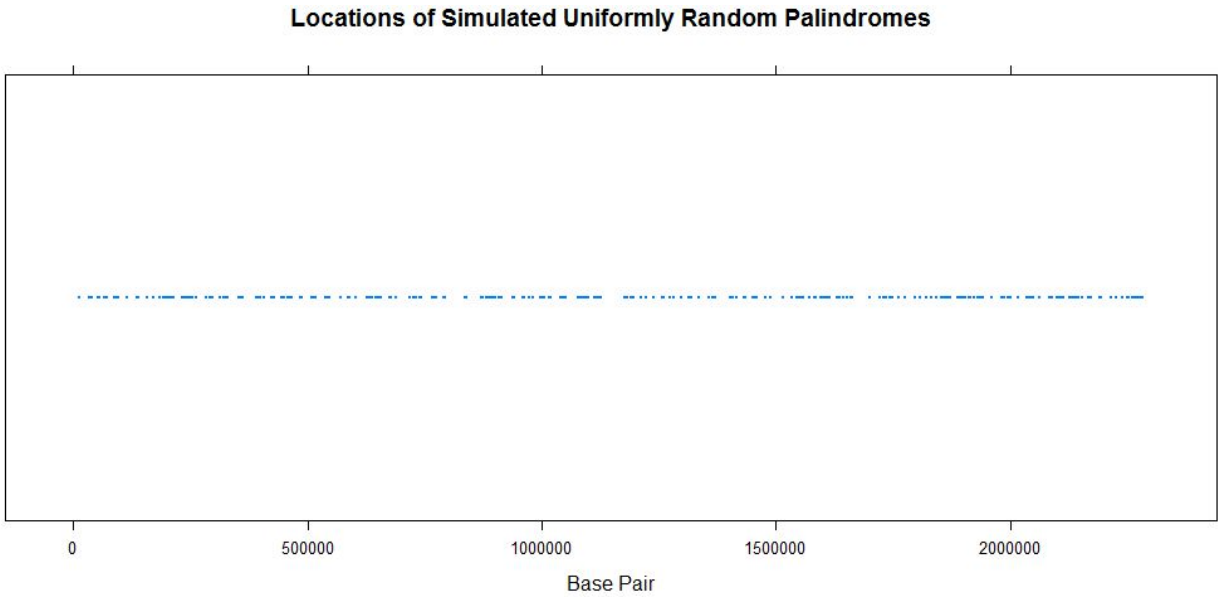
By binning our data, we can observe the counts of the palindromes in order to better quantify how normal or abnormal the DNA sequence is. In Figure 13 below, we can see the observed sequence of palindromes and its various clusters along the chosen fifty-seven bins of the approximate length of 4,000 base pairs.

Figure 13: Palindrome Locations



To further see any discrepancies between the observed sequence and the expected, we can simulate data based off of the uniform distribution. Below is the locational data of the palindromes that are uniform randomly scattered. Although it is hard to discern, our observed DNA sequence has longer and thicker clusters of palindromes when they do occur. The simulated data however, has nice segments of empty spaces and consistent clusters.

Figure 14: Simulated Palindrome Locations



We continue to use this binning technique to realize the counts of palindromes for each bin, which can be seen below.

Table 1: Comparison of Count of Clusters

Bin	Observed	Simulated
4023.754	7	3
8047.509	1	7
12071.26	5	6
16095.02	4	4
.....		
221306.5	1	3
225330.2	8	6
229354	7	6

Now that the frequency of palindromes are quantified we can see how our observed data fits the uniform or “random scatter” sequence we simulated. We can achieve this through the Chi-Square goodness of fit test. As mentioned earlier in the

paper, our null hypothesis is that our observed palindrome locations is consistent to the uniform distribution. A low p-value from the chi-square test results in us rejecting the null hypothesis in favor of the alternative: The observed sequence does not follow the uniform distribution. In fact, we do reject the null hypothesis with a p-value of .01937.

We can also use the chi-square test in order to determine the type of distribution of the given data. The null hypothesis for the chi-square test assumed to be a Poisson distribution. When we assume 60 intervals of equal size between 1 and 229354, we obtain table with the count range, the actual counts and the expected counts for a Poisson distribution as,

Table 2: Observed vs. Expected Palindrome counts within Intervals

Palindrome count	No. of intervals (Observed)	No. of intervals (Expected)
0-2	10	7.823
3	9	8.648
4	9	10.666
5	10	10.523
6	5	8.652
7	10	6.098
8+	7	7.554

We obtain the pearson's chi-squared value to be 4.986. On looking up the value from the chi-square table, we get a p-value of 0.418. This tells us that we fail to reject the null hypothesis of the chi-square test with a confidence interval of 90%, which is the given data is a Poisson distribution.

4. The biggest cluster

For the last part of our data analysis, we attempted to identify if any of the DNA palindromic clusters represented the CMV origin of replication. For a potential CMV origin of replication, we expect to see a large number of palindromes clustered in close proximity.

Figure 15: Distribution of location of palindromes in CMV DNA.

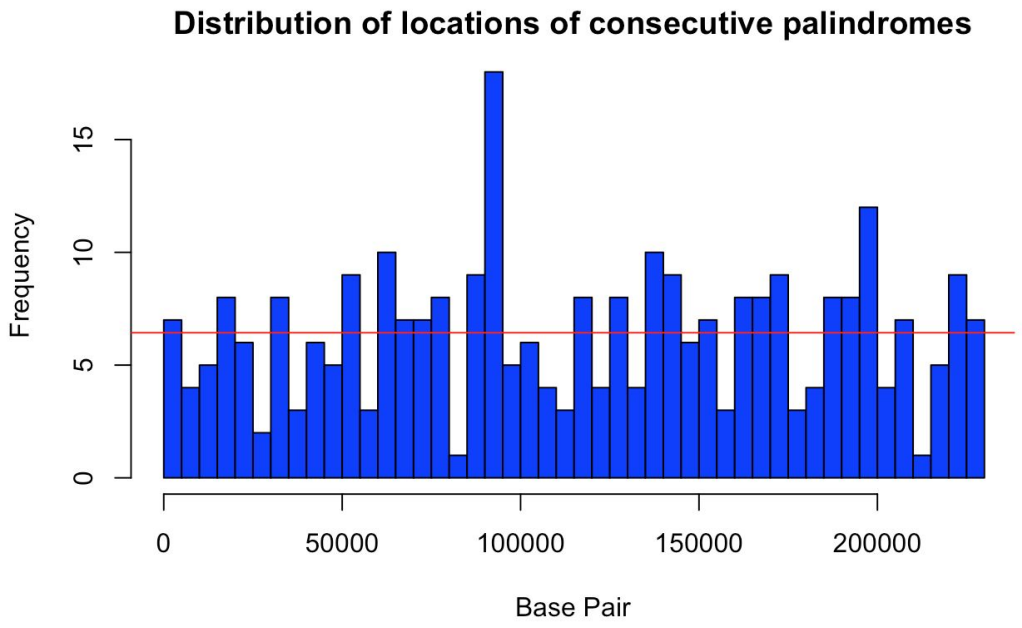


Figure 16: Histogram of the number of palindromes per 5Kbp in CMV DNA.

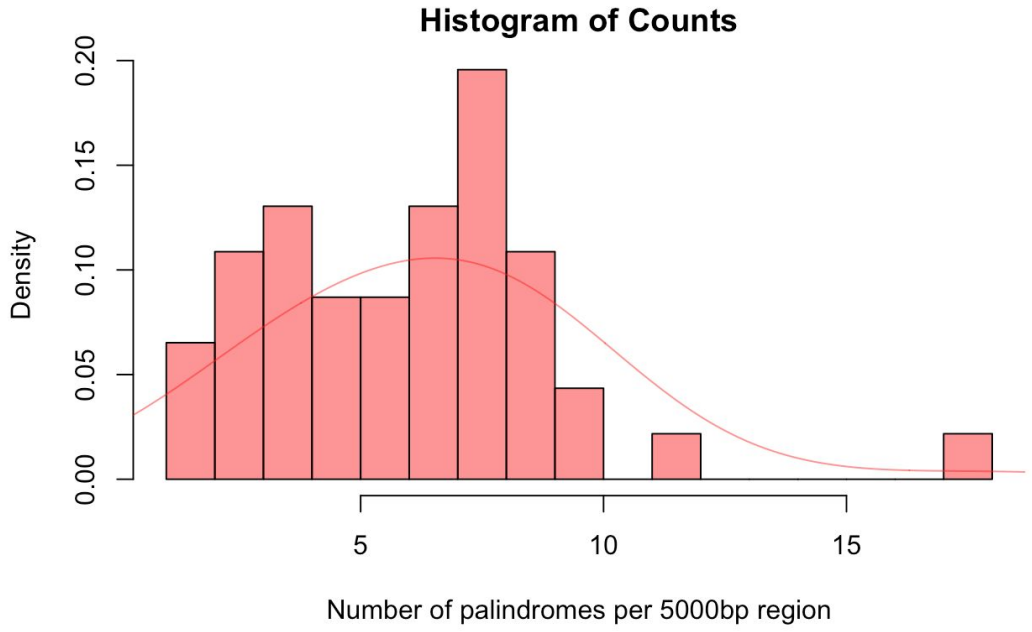
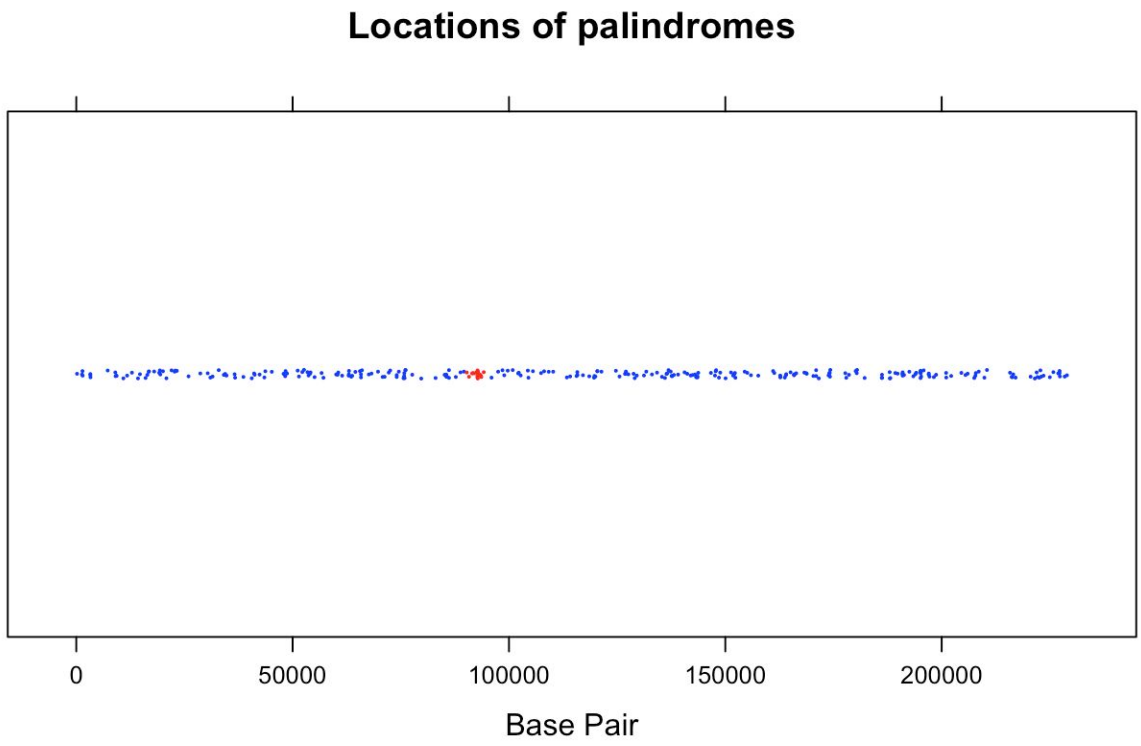


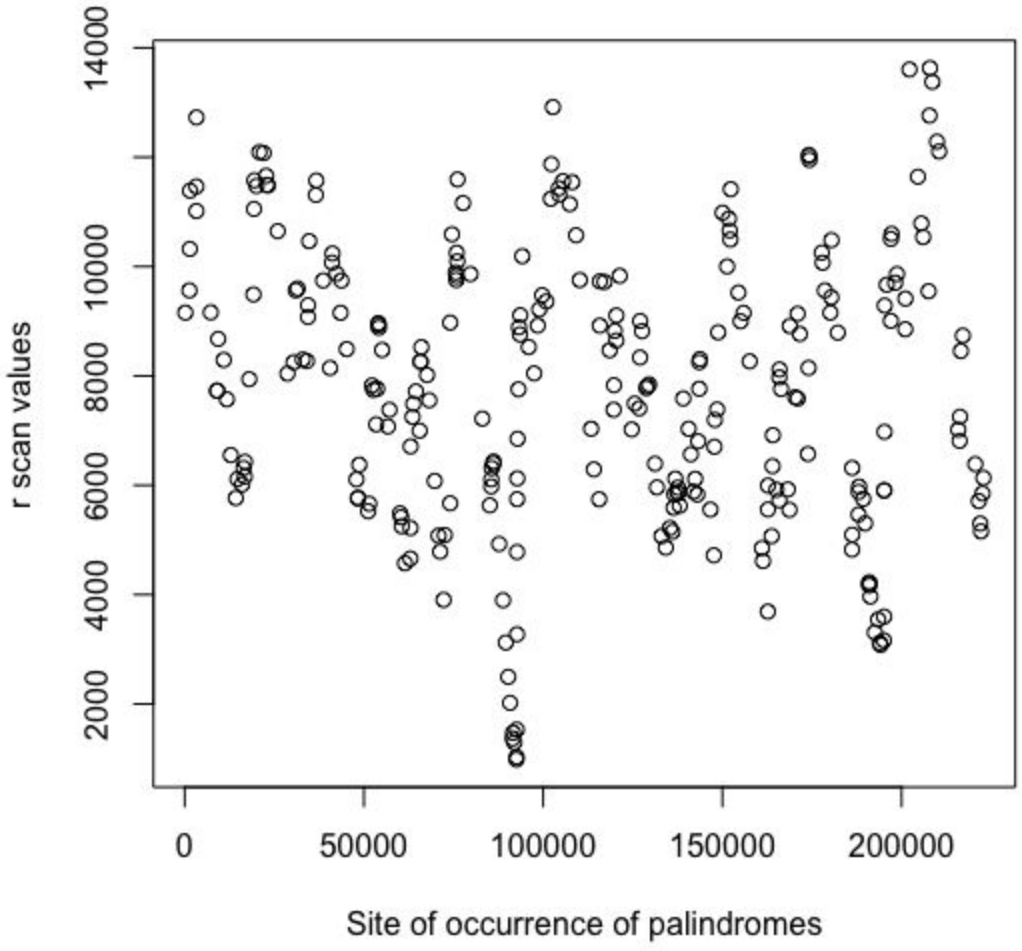
Figure 17: Locations of palindromes in CMV DNA.



In the previous section (3. Counts), we analyzed the number of palindromes in various non-overlapping segments of the DNA. After looking at multiple bin widths, we found 5000bp intervals to be most sensitive for searching for the origin of replication. We chose this size (5000bp) interval because it was not too small and therefore not missing large clusters of palindromes, and the interval was not too large to miss small deviations from random scatter. Figure 15 shows the distribution of the base pair location of palindromes (blue), and the red line is at the frequency 6.435, which is the mean number of counts per 5000 bp. We hypothesized this distribution to follow a uniform distribution if the occurrence of palindromes are truly random. As described previously, we performed a chi-square goodness of fit test. To do this we used the counts of palindromes per 5000bp bins in the R chisq.test function, and concluded to reject this hypothesis at a significance level of alpha=0.05 (χ^2 = 66.716, df = 45, p-value = 0.01937). Figure 16 shows the number of palindromes in non-overlapping segments of 5000 base pairs. We expect this distribution to follow a Poisson distribution if the count of palindromes are truly random. As described previously, we performed a chi-square test and conclude to reject this hypothesis.

Seeing as the chi-square tests result in us rejecting the hypothesis that the location of palindromes is random, we can assume the occurrence of a cluster of palindrome is due to a biological phenomena.We found the region 90,000-95,000 (red) in the DNA to contain the greatest number of palindromes (18 palindromes) (Figure 17). We can assume this region to be the CMV DNA origin of replication because 18 palindromes in this region is not likely to occur by random chance.

Figure 18: Plot of the site of occurrence vs the r scan values



We also used the r-scan statistics as an additional method to find significant clusters of importance in determining the origin of replication of the CMV. We take the difference between the r consecutive locations and add them together. If the value obtained is low, it signifies the occurrence of many consecutive palindromes. From Figure 18, we can observe that there is an unusual cluster at the location 92,526. This additional method supports the previous analysis in selecting the 90,000-95,000 bp DNA region to be the CMV origin of replication.

CONCLUSION

We statistically analyzed the locations of palindromes on a CMV DNA sequence in an attempt to find a potential origin of replication site. We compared our observed data to random distributions, and performed multiple statistical hypothesis tests to determine if the palindromic clusters were or were not by chance. Out null hypothesis was that the palindromes are randomly scattered. Our hypothesis tests concluded that we could reject the null hypothesis. In Section 3, our tests suggested that the data does not follow a uniform distribution, but we did prove to a 90% confidence level that the appearance of palindromes fits a Poisson distribution. In Section 4, tests suggested that it is highly unlikely for the 18 palindromes (the largest found cluster) to be found between locations 90,000-95,000 by chance. Additionally, the high volume of consecutive palindromes around location 92,526 further suggests the origin of replication is in that region. Based on our rigorous statistical analysis, we recommend biologists investigate the DNA region 90,000-95,000 bp for the CMV origin of replication. It would be highly cost effective for biological researchers to experiment on this location first seeing as our analysis has determined the clusters of palindromes in this region are not likely to occur by chance.

REFERENCES

Ahmed, A. "Antiviral Treatment of Cytomegalovirus Infection." *Infectious Disorders Drug*

Targets. U.S. National Library of Medicine, Oct. 2011. Web. 22 Feb. 2017.

https://www.ncbi.nlm.nih.gov/pubmed/21827432

Chee, M.S. et al. “The DNA sequence of the human cytomegalovirus genome.” *DNA Sequence* 2:1 (1991): pages unknown. Web. Accessed 22 Feb 2017. http://www.tandfonline.com/toc/imdn21/2/1?quickLinkVolume=2&quickLinkIssue =1&selectedTab=citation&volume=2

“Cytomegalovirus (CMV) and Congenital CMV Infection”. *CDC*, 17 June 2016, https://www.cdc.gov/cmv/overview.html. Accessed 13 Feb 2017

Giel-Pietraszuk, M., M. Hoffmann, S. Dolecka, J. Rychlewski, and J. Barciszewski. "Palindromes in Proteins." *Journal of Protein Chemistry*. U.S. National Library of Medicine, 22 Feb. 2003. Web. Accessed 14 Feb. 2017. https://www.ncbi.nlm.nih.gov/pubmed/12760415

Leung, Ming Ying et al. “An efficient algorithm for identifying matches with errors in multiple long molecular sequences.” *Journal of Molecular Biology* 221.12 (1991): 1376-1378. Web. 22 Feb 2017. https://utep.influuent.utsystem.edu/en/publications/an-efficient-algorithm-for-identifying-matches-with-errors-in-mul

Masse, M. J., S. Karlin, G. A. Schachtel, and E. S. Mocarski. "Human Cytomegalovirus Origin of DNA Replication (oriLyt) Resides within a Highly Complex Repetitive Region." Proceedings of the National Academy of Sciences 89.12 (1992): 5246-250. Web. 14 Feb. 2017. http://www.pnas.org/content/89/12/5246.full.pdf

Rosa, Corinna La, and Don J. Diamond. "The Immune Response to Human CMV." *Future Virology*. U.S. National Library of Medicine, 01 Mar. 2012. Web. 14 Feb. 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539762/

Swanson, Elizabeth C., and Mark R. Schleiss. "Congenital Cytomegalovirus Infection:

New

Prospects for Prevention and Therapy: for Pediatric Clinics of North America:

Advances in Evaluation, Diagnosis and Treatment of Pediatric Infectious Disease."

Pediatric Clinics of North America. U.S. National Library of Medicine, Apr. 2013.

Web. 14 Feb. 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807860/

R Scan Statistics: http://www.cmbl.uga.edu/software/r-scans.htm