Case Study 3:
Patterns in DNA Analysis

Caitlyn Bryant - Applied Mathematics, 4th year
Edward Chao - Mathematics: Statistics, 4th year
Kieran Mann - Mathematics/Economics, 4th year
Matthew Kucirek - Applied Mathematics, 4th year

Table of Contents

Introduction

In order to find ways to fight the human cytomegalovirus (a potentially life-threatening disease for people with a suppressed immune system), scientists must study the way in which the virus replicates. They look for the origin of replication on the virus' DNA, which is flagged by patterns found in the DNA. One of these types of patterns that leads to the origin of replication on the DNA is the complementary palindrome pattern. Our objective is to find unusual clusters of complementary palindromes, which will help us narrow the search in finding the origin of replication.

When searching for unusually dense clusters of palindromes, we will: (1) determine whether the given data set of palindrome locations is random by comparing it to a random scatter, (2) use graphical methods to examine the spacings between consecutive palindromes and compare locations of the palindromes, (3) use graphical methods to examine the counts of palindromes in various regions of the DNA, (4) determine whether or not the interval with the greatest number of palindromes indicates a potential origin of replication.

The primary data that we studied consisted of the DNA sequence of the human cytomegalovirus (CMV) that was published in 1990 and is 229,354 letters long. In 1991, Leung et al. implemented search algorithms in order to search the sequence for several types of patterns. 296 palindromes were found that were at least 10 letters long, and of these 296 palindromes, the longest ones were found to be 18 letters long occurring in locations: 14719, 75812, 90763, and 173893. Occurrences of palindromes shorter than 10 letters were ignored.

The "big question" that we will answer by our data analyses is: What advise would we give to a biologist who is about to start experimentally searching for the origin of replication?

In the remainder of our paper, we will summarize our research in all four of our categories of analysis and use statistical methods to come up with advice to give to a biologist working on finding the virus' DNA's origin of replication.

Research and Analysis

2.1. Random Scatter

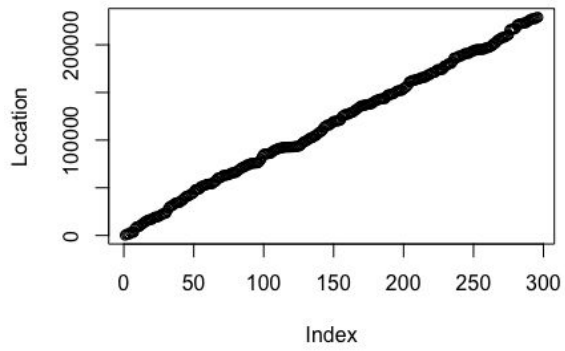

Figure 1: *A scatterplot of our given palindrome location data*

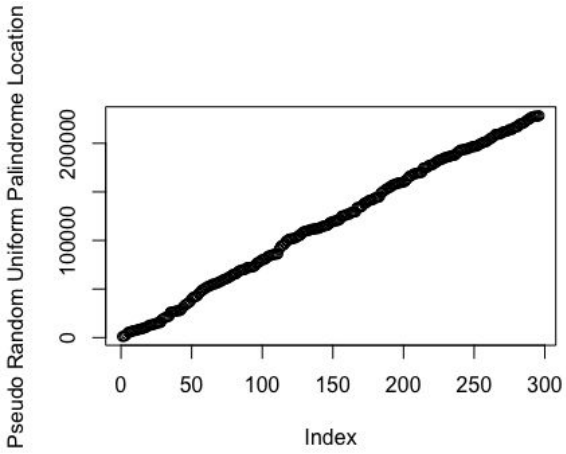

Figure 2: *A scatterplot of the pseudo random uniform palindrome location data*

We have data that consists of 296 palindrome locations in a DNA sequence of 229,354 bases. In order to determine if there is a pattern in this data or if they are a random occurrence, we will compare this to a random uniform distribution of 296 numbers between 1 and 229,354. We generated these numbers using R's pseudo random number generator.

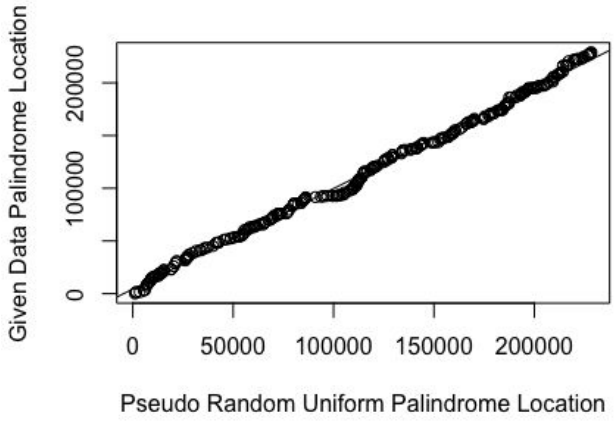

Figure 3: *A scatterplot of the pseudo random uniform palindrome locations and our given data palindrome locations with a linear regression line*

We made a scatterplot using the pseudo random uniform location data set on the x-axis and our given palindrome location data set on the y-axis. We then added a linear regression line and calculated R^2. Since R^2=0.9941033 and this value is very close to 1, this proves that the palindrome data is, in fact, random. In order to ensure that this conclusion is true, we found R^2 100 different times with 100 different sets of 296 pseudo random uniform palindrome locations. We then made a histogram of this data and analyzed the R^2 values further.

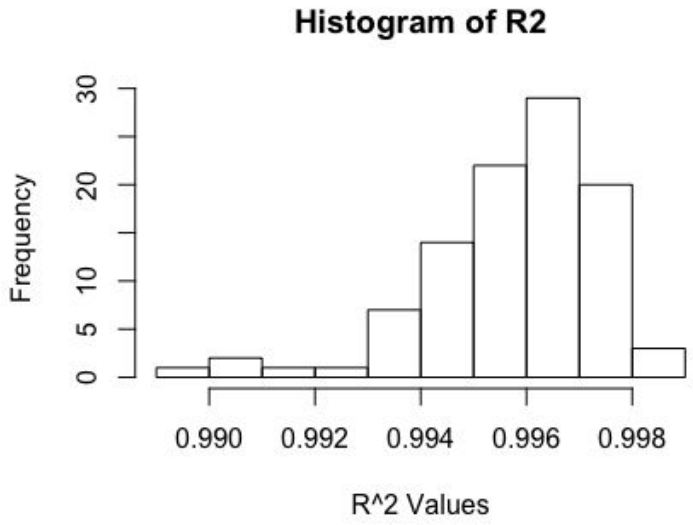

Figure 4: *A histogram of 100 different R^2 values calculated from 100 different data sets of 296 pseudo random uniform palindrome locations*

Mean	0.9957758
Standard deviation	0.001659481
Median	0.9960345
Maximum	0.9984117
Minimum	0.9896372
Skewness	-1.254055
Kurtosis	5.161482
Quantile	0% 25% 50% 75% 100% 0.9896372 0.9949278 0.9960345 0.9969056 0.9984117

The analysis of our R^2 values show that R^2 is approximately equal to 1, which leads us to conclude that our given palindrome location data set is random. Although we now know the palindromes' locations are random, we are still unsure of how exactly the locations are distributed; we will analyze this in the next section.

2.2. Locations and Spacings

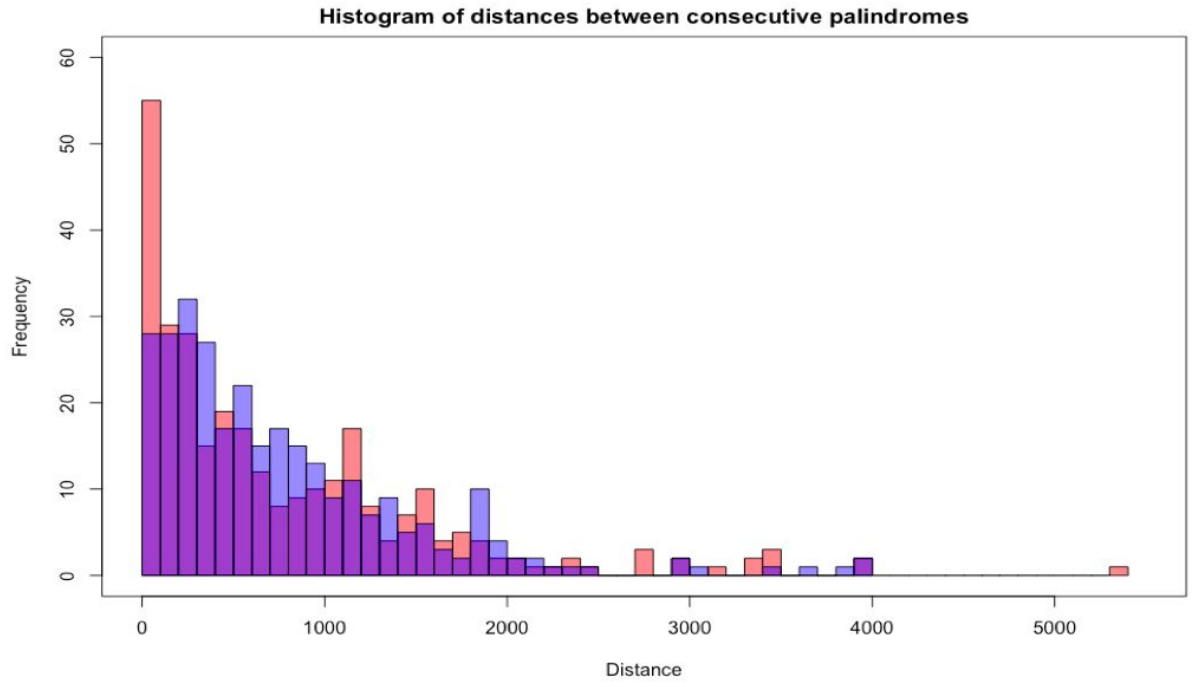

Figure 2.2.1: Histogram of observed DNA (red) consecutive palindrome distances versus a uniform random sample's (blue) consecutive palindrome distances

In examining the distances between the pairings looking for any irregularities, we compared the observed distribution to a pseudo-random sample generated from selecting 296 indices for palindrome locations at random from a sequence of 1 to 229,354 potential locations distributed by a uniform random distribution. We created an algorithm that applies a simple difference function on a rolling interval of 2, moving by 1 index each time. This measures the distance between consecutive base pairs and allows us to search for patterns in the distribution of via the proxy distances between palindromes.

Looking at the histogram created by examining such distances between consecutive palindrome, it is clear that far more palindromes than one would expect to observe out of a truly random list of palindrome locations occur quite near each other. In applying a Kormogorov-Smirnov as a method of taking a non-parametric deviance of the two distributions, a p-value of 0.1156 was yielded in our draw. Visually analyzing the distribution, it appears to be a mean-preserving spread of some sort of a uniform population, but it is heavily weighted toward certain points in the distributions, specifically close

palindromes and ones occurring about 1,000-1,200 segments away, or at a few other central clustering points.

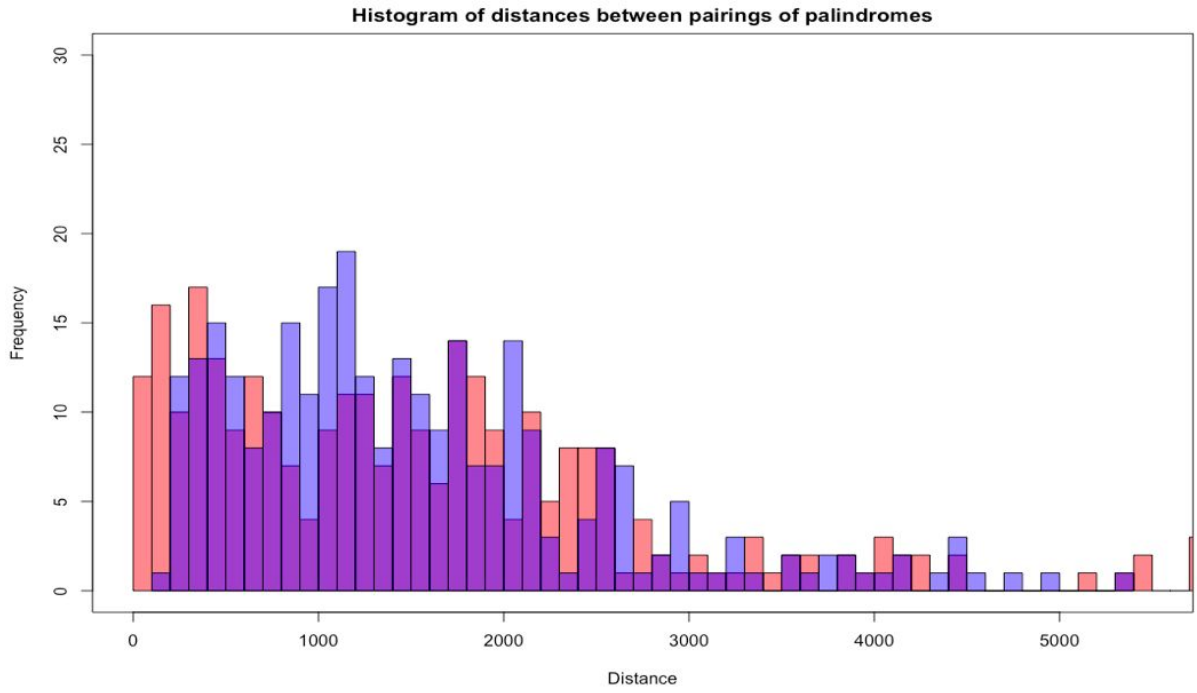

Figure 2.2.2: Overlayed histograms of observed (red) and random (blue) distances between palindromes

In looking at the distances between each consecutive pairing of palindromes (no two occur directly adjacent to one another) against the same two distributions, a pattern quickly becomes more evident. The pairs of palindromes that are either quite close to one another or about 2,000 segments away are quite unlikely. In addition, it seems there are a good number of pairs of palindromes that are quite a few segments away from one another, but it is certainly harder to make a significant conclusion about such a small spike. Applying a Kormogorov-Smirnov test to the full distributions, a larger p-value of 0.4391 is yielded, signalling an even more significant departure from a random normal than the distribution of distances between consecutive palindromes than when they're in pairs.

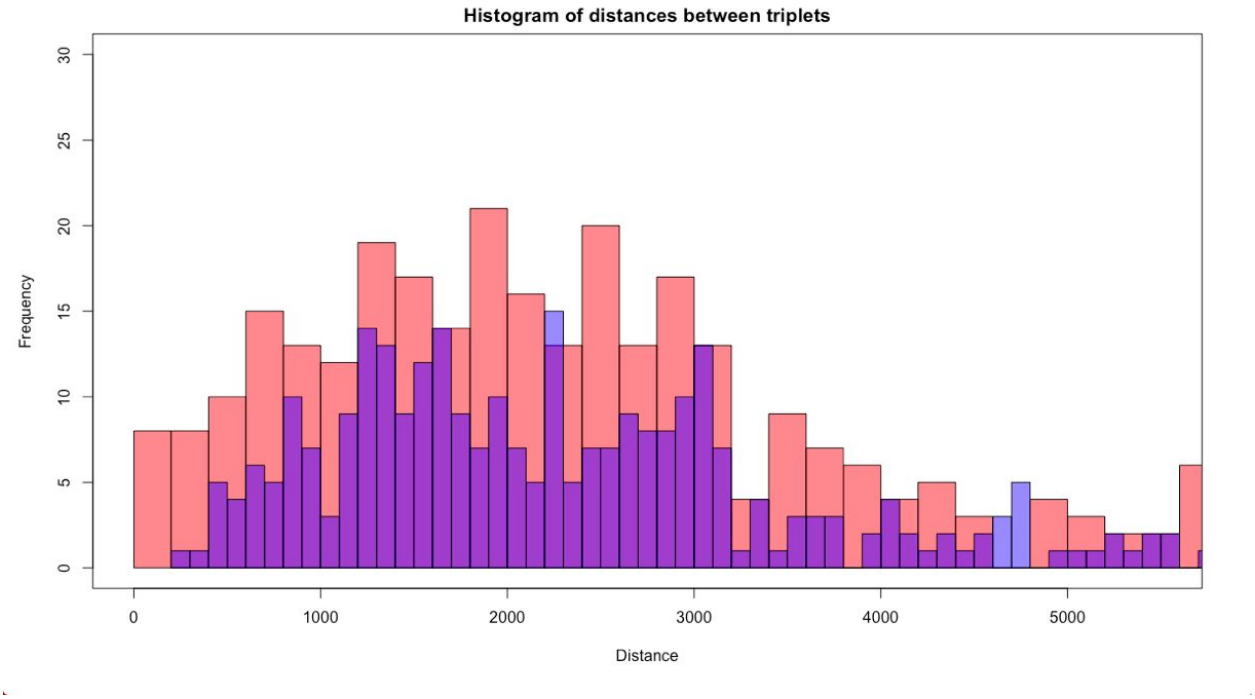

Figure 2.2.3: Taking it up one more notch, a histogram of the distribution of distances between observed palindromes (red) and a theoretical random sample (blue)

Going one further and examining the distances between each grouping of three palindromes yields and even more significant departure from the distances one would expect to observe from a randomly distributed DNA sequence. The skew toward very close groups of three consecutive palindromes is highly significant, both visually on a histogram and probably on any number of more formal statistical tests. For one, a Kormogorov-Smirnov test with a two-sided null hypothesis yields a p-value of 0.9336, meaning that is quite unlikely (to say the least) that the observed DNA is uniformly randomly distributed. It seems that scientists looking for the origin of replication should definitely consider the distances between triplets of palindromes in their search for the origin of replication.

2.3. Counts

By splitting the DNA into equal, non-overlapping regions, we were able to better compare the number of palindromes located in an interval to the expected number of a uniform random scatter.

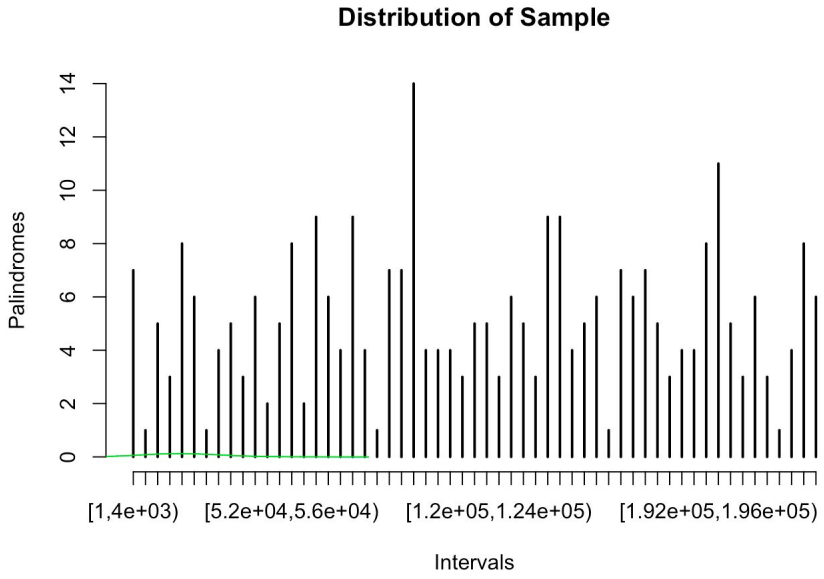

Figure 1: *Distribution of palindrome locations*

On a whole, it is a reasonable assumption that the distribution of the samples would be in line with normality, but when the locations are divided into intervals and plotted based on the count of palindromes, the distribution is far from being normal. In Figure 1, the density line is far skewed to the left with only a marginal curve around the interval [4004,8008], with no existence on the rest of the graph. This leads us to believe that the distribution of the intervals would follow a Poisson Distribution. So to determine if the Poisson distribution reasonably fits the data, we conducted a goodness of fit test. We began first with a graph of the distribution of the number of palindrome counts in an interval.

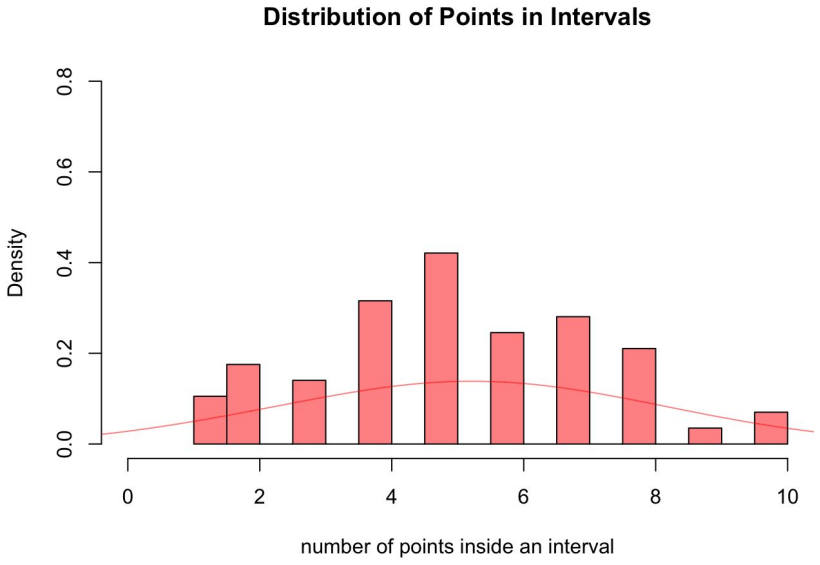

Figure 2a: *Histogram of the distribution of palindrome location, points, inside an interval*

On its own, the graph of the distribution of palindrome locations, or points, follows what we would expect a Poisson distribution to look like. In Figure 2a, the points between 4 and 6 make up most of the distribution, suggesting that due to the vast number of palindromes, it became likely that some of the palindromes would cluster around the same interval. And after a comparison of distributions between our sample and a random scatter, our assumptions became confirmed.

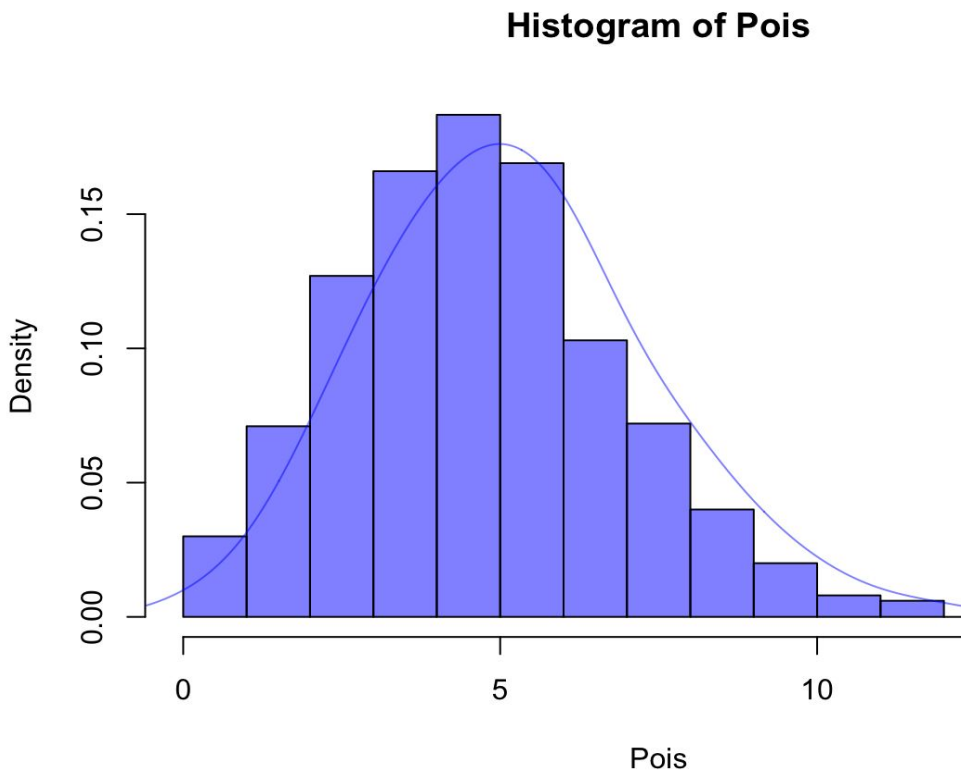

Figure 2b: *The distribution of the random scatter*

In Figure 2b, the area around 4-6 make up the majority of the graph itself with the peak of the curve at 4. The density line in Figure 2a follows the one pictured in Figure 2b, giving way to the idea that the data fits reasonably well in line with the Poisson distribution. Due to the nature of this distribution, we can safely assume not only that the probability of a palindrome to fall in a particular interval is independent of the other palindromes around it, but also that no two palindromes fell in the same place.

But in order to validate our claims, we did further tests. We propose that if the random scatter holds true, then we can calculate the chi-square distribution to measure how well the distribution fits with the data. Using 57 intervals and 296 samples, we calculated our χ^2 to be 52.9. We used this value to compute the probability of having another test statistic of similar size, which is 0.999. This large number suggests that deviations are extremely likely, affirming the Poisson distribution of our palindrome locations. And because our value was so large, we did not feel the need to conduct a standardized residual on our distributions because we did not observe a lack of fit.

2.4. The Biggest Cluster

In this section we aim to determine if the interval with the largest number of palindromes is statistically significant compared to the other intervals. One thing we are concerned with is the interval length that we will choose in order to partition the DNA sequence. A length that is too small will cut off meaningful clusters of palindromes and a length that is too large will add noise to the data, overcompensating for clusters. The lengths we tested were 2000, 3000, 4000 and 5000. With an interval length of 2000 we observe that there are 10 intervals which contain zero palindromes and 25 intervals that contain only one palindrome. This means that roughly 30% of the intervals do not contain a cluster. Looking at a length of 3000 we see that nearly 16% of intervals contain either zero or one palindromes and looking at a length of 4000 we see that roughly 9% of intervals have either zero or one palindrome. Finally, looking at interval lengths of 5000 we observe that 4.4% of intervals contain either zero or one palindrome. From this we see that we want to either look at interval lengths of 4000 or 5000.
The next thing we want to determine is the distribution of palindrome locations throughout the DNA. First, we compare our location data to a uniform distribution. To obtain our distribution, we utilized R's runif function to randomly generate 296 locations between 1 and 229345. We chose 229345 because the minimum palindrome length is 10, thus the last possible base pair that could start a palindrome would be in location 229345 (of course this assumes there are no palindromes within palindromes).

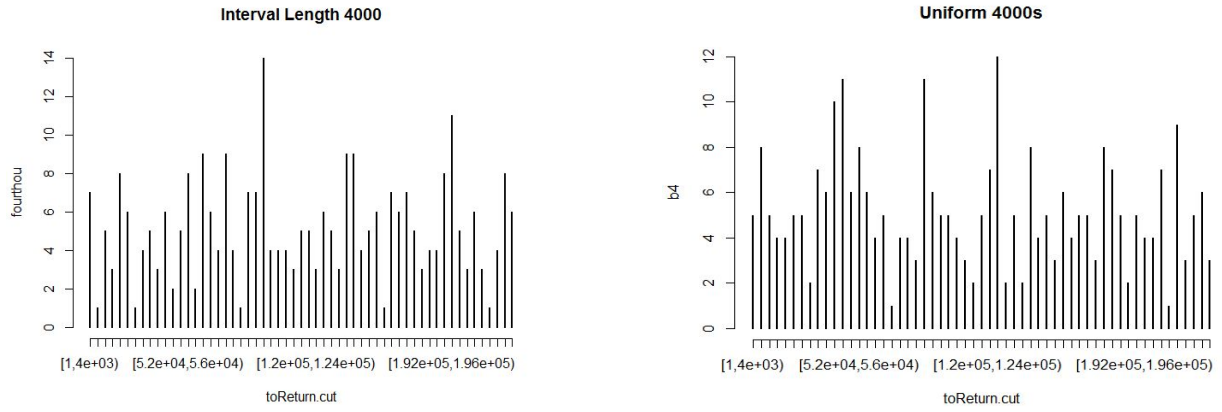

Figure 2.4.1 Comparing our data in lengths of 4000 to uniform data

At a first glance, the two distributions appear similar. They both achieve their maximum near location 100000. Moreover, the distributions start to behave similarly near the 175000 location and continue to do so for the rest of the DNA strand. However, from the beginning of the DNA until the maximum, the distributions behavior varies. Our data has nine or fewer palindromes in every interval until the maximum interval. In contrast, the uniform scatter achieves values larger than nine many times before reaching the maximum interval. We also took a look at distributions over interval lengths of 5000 and saw very similar results, where up until the maximum the distributions behave differently and towards the last half they perform consistently with one another.

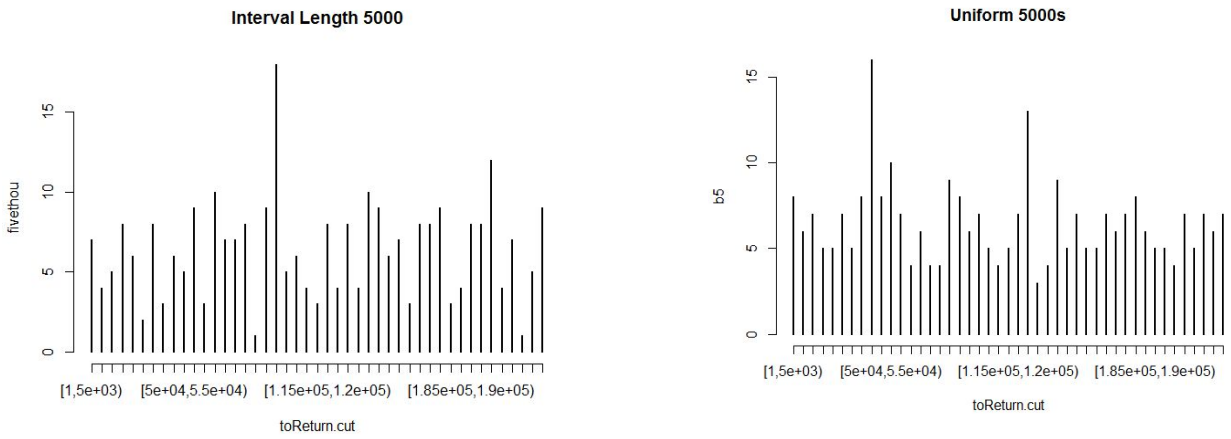

Figure 2.4.2 Comparing our data in lengths of 5000 to uniform data

With this, we determine that a uniform distribution is not the best approximation for our location distribution. Our next guess is that the data will fit a Poisson distribution. Here we still carry out our calculations for both interval lengths of 4000 and 5000. For interval lengths of 4000 we have lambda valued at 5.19 and for interval lengths of 5000 we have lambda valued at 6.58. Lambda is the mean number of palindromes located in an interval. To see whether a Poisson is good fit, we use R to generate 296 locations using the rpois function, graphed the histogram and kernel density estimation, and compared it to the histogram and density of our data.

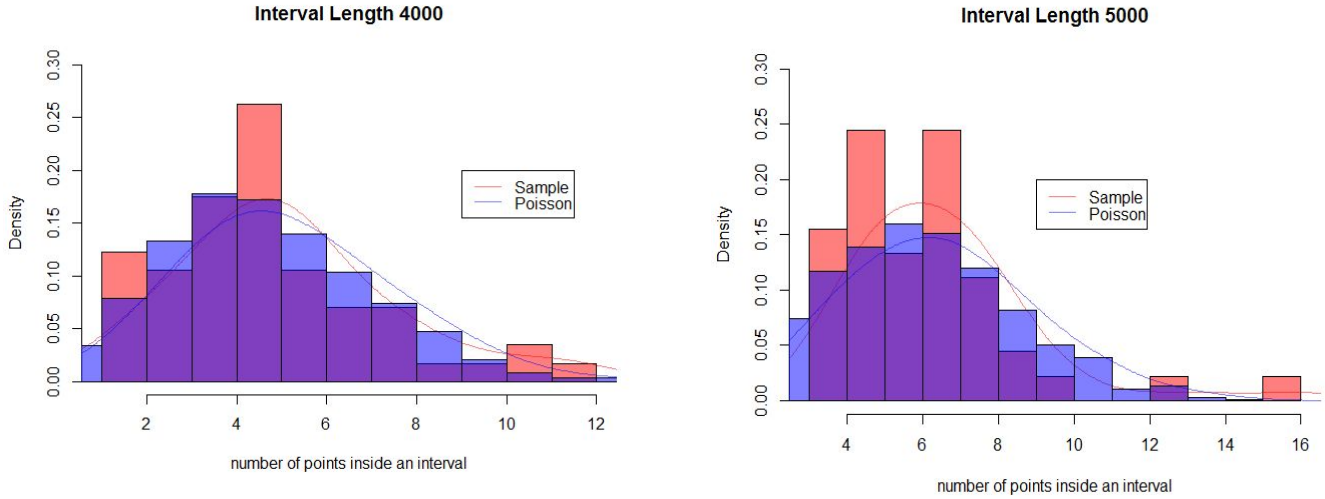

Figure 2.4.3 Comparing our data to a Poisson distribution

Visually, our data appears to coincide with a Poisson distribution. In order to validate this, we run a Chi-Squared goodness of fit test. Before running the test we look at the observed number of intervals that contain 0,1,2,… palindromes. Next, we use the Poisson model to generate the expected number of intervals that contain 0,1,2,… palindromes. When calculating the expected number of intervals for certain

values, we see that we get results such as 0.5 intervals or 1.2 intervals. In order to make more sense of the data and get a better test statistic, we group intervals together so that the minimum number of intervals for a particular number of palindromes is 3. Now that we have our observed and expected values, we can calculate our test statistic and run the Chi-Squared test. Both tests return high p-values meaning that we fail to reject the null hypothesis and determine that a Poisson distribution is in fact a good fit. For interval lengths of 4000 we obtain a p-value of 0.98 and with interval lengths of 5000 we obtain a p-value of 0.22. This indicates that using an interval length of 4000 is optimal.

Now that we have a fitted distribution, we analyze the distribution of the maximum. Our lambda value is greater than 5, so we begin our initial guessing of the maximum is 5. Our global maximum is 14 so that is where our search will stop. Running through these values, we find the probability that the maximum could occur. Doing this, we see that the probability that the maximum number of palindromes in any interval equals or exceeds 14 is 0.0189, which is small enough to consider this maximum statistically significant.

Conclusion

The conclusions we draw from our analyses are: (1) our given palindrome location data set is random, (2), the distances between pairings or triplets of palindromes is highly unlikely to be distributed by a uniform random distribution, (3) the probability that a palindrome will fall in a particular interval is independent of the other palindromes around it, and no two palindromes fell in the same place, (4) from seeing that there are relatively few palindromes in any interval up to the maximum interval and seeing that the maximum interval is statistically significant we conclude that the maximum interval is a good starting place in the search for the replication site.

The advice we would give a biologist who is about to start experimentally searching for the origin of replication in the virus' DNA is to compare the given data to multiple data sets of random scatter in order to detect if the cluster of palindromes occurred by chance or if it is something to pay more attention to. If, after repeated simulations of random data sets, you deduce that the cluster is highly unlikely to have occurred by chance, consider this region to be a potential replication site and investigate it further. Looking at the distances between consecutive pairings or triplets of palindromes seems to yield a pattern, so searching at regular intervals and unusually close to each palindrome may cut down time and cost on experiments.

Appendices

Theory

1. Poisson Distribution

We divided each 4000 base pairs into 57 sub-intervals, to count how many data points, or palindrome locations, would appear. We believe that the number of points would follow a Poisson distribution, and the rate is equal to the expected number of points in an interval.

2. Goodness of Fit

We used Pearson's χ^2 test to determine how well the data fit the distribution. The test statistic was calculated by normalizing the sum of deviations between the observed and theoretical observations. Then, using 56 degrees of freedom, we were able to calculate the p-value. A large p-value would signify a goodness of fit, whereas a small value would've signaled a lack of one.

3. Maximum of a random variable

We calculated the distribution for the maximum number of palindromes in any interval. Since the intervals are disjoint and the number of palindromes in nonoverlapping intervals are independent, we simply take the probability density for one interval, and raise it to the power of the number of intervals we have. In this case, we had 57 intervals, so the distribution for the maximum is simply the poisson odf raised to the 57th power

In determining the deviance of the distribution of distances between consecutive, pairs and triplets of palindromes, we use the Kolmogorov-Smirnov (KS) test, a nonparametric test for describing cumulative deviance of the distributions behind two samples. The KS test is convenient to apply when looking at differences in distributions that may have similar medians but hold their densities differently.

Code

(See attached r markdown file for replicable code.)