Shawn Koohy

3 April 2022

Dr. Davis

MTH332

DNA Replication, Palindromes, and Their Relation to Probabilistic Distributions

Abstract

In the scientific study of genes, genetics, DNA replication is an important aspect to the growth, development, and renewal of cells. It is important to understand where the DNA replication occurs or rather how the DNA replication is distributed along a defined length of DNA. We will be discussing the locations of what are known as palindromes, how they are connected to DNA, how they are distributed along DNA, and their possible involvement in the replication of DNA. To do this, probabilistic measurements and statistical tests will be implemented to investigate the randomness of these palindrome locations. Are the palindromes distributed along DNA uniformly, follow a Poisson distribution or some other form of a probabilistic model. It will be discovered that the location distribution can have a differing conclusion based on the perspective. Our analysis and examination lead us to consider more likely than not that the Palindromes may be distributed uniformly if we were removing the outliers found in our data. The reason why we would want to remove those outliers is because we may believe those are the locations of replication (non-random). However, anther conclusion is that if we follow our implemented statistical test, the χ^2 test, we will believe that the palindromes follow a Poisson distribution.

Introduction & Background

DNA is a self-replicating material that exists in most living organisms and is known to be the carrier of genetic information that effects the development, functionality, growth, and reproduction of cells. In a more descriptive case, DNA is a long-coded message made of four letters: A, C, G and T. Since this representation of DNA contains only four letters, the possibility of seeing sequences of patterns is very high. It is hypothesized that these patterns may be the site of origin for DNA replication. In genetics or the study of DNA, A is considered to be complementary to T, and G is complementary to C. One of the type of patterns that is common in DNA is palindromes or more specifically complementary palindromes. A palindrome is a word, sentence or sequence of letters that is the same backwards that it is forwards. Some examples of these are "madam", "race car", "taco cat" or "ABCDEDCBA". When it comes to complementary palindromes, let's look an example of one. Consider the DNA sequence "GGGCATGCCC". When we first look at this, we may not see any palindromes in sight because "GGGCATGCCC" is not the same as "CCCGTACGGG". When we consider complimentary palindromes such that A complement is T and G complement is C we get "GGGCATGCCC" (forwards) and "CCCGTACGGG" (backwards) $\xrightarrow{yields}$ "GGGCATGCCC". This means that "GGGCATGCCC" is a complementary palindrome. The data given for the report is a set of locations for which DNA sequences contain palindromes of length 10 or more. For example, in the data set the first point is 177. This means that in position/location 177 there is a palindrome of length 10 or more. We will be considering a data set containing 294 palindromes of length 10 or more. Our next point in the data set is 1321 and our final point will be 227,316 (All locations between 0 - 228,000). The goal of this research is to determine or address the issue on how these locations are distributed, are they uniformly random, follow a poison distribution, or maybe follow some other type of distribution. When we consider something to be uniformly random, we will take this as saying that if we split the totally length of the DNA in n equal segments then each of those segments will contain the same number of palindromes. If we split the DNA into 50 equal segments, we expect the same number of palindromes in the first segment, the second and so on. When it comes to how we define a Poisson distribution or more correctly a homogeneous Poisson process there are three main assumptions to be accepted or that must be satisfied. The first one being that the λ (the expect number of events in the interval) does not change with

location, each segment will have the same λ. The next two assumptions are that the location of the points are independent of each other and no two points overlap or are in the same exact place. The Poisson distribution, or probability at a value k is given by,

$$P(K = k) = \frac{\lambda^k e^{-\lambda}}{k!}, for\ k = 0, 1, 2, \ldots$$

Methods

We will start by importing the data of complementary palindrome locations into Mathematica. For simplicity we will call complementary palindromes just palindromes. Once we have our locations, we can start with a probability density function histogram. This will show us the uniformity of our data. In our analysis we will be splitting the total length of DNA into 4 different cases. In the first case we will split the total DNA into 50 equal lengthen segments of 4560, because $4560 * 50 = 228000$. We should also reinforce the idea that we are looking at the first 294 palindromes of the data which is contained in the range $[0, 280{,}000]$. The next 3 case we split the DNA into 57, 50 and 65 equal length segments. Each segment will have a length of 4000, 3800, and 3508, respectively. Other than the PDF histogram we will be creating histograms of different bin sizes. This bin sizes will correspond with the number of equal lengthened segments that we split the data up as. As an example, for the 65-segment case we will create a histogram with 65 bins. This will allow us to see the distribution of palindromes for more specific cases of data splitting and for more specific regions in the data. This is also done to possibly identify a sight of DNA replication, a secondary goal of this report. We will be finding the descriptive statistics of the location to better understand how our data behaves. Once this is all completed, we will apply a computational test to see total number of palindromes in each segment that we have created. We are going to be grouping the number of palindromes for each interval/segment observed. We will see how many segments contain 0-2, 3, 4,…,8,9+ palindromes. For example, in the 57-segment case there were 7 intervals that contains 0-2 palindromes and 8 intervals that contained 6 palindromes. Similarly, we will need an expected number of intervals. This is where we will start gathering data on uniform randomness and Poisson distributions. The case of uniform randomly distributed positions is very simply to calculate. If we have a segment of length 57, we expect the number of palindromes in each

segment to be 5.16 or $\cong \frac{294}{57}$. This will be our expected number of intervals containing x number of palindromes. Next, we need to see how to find the expected values for our Poisson distribution. We will first define our λ as $\frac{Number\ of\ Palindromes}{Number\ of\ Intervals}$ or just $\frac{294}{Number\ of\ Intervals}$. Then we will test $k = 0, 1, 2$ into,

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

and sum them together. The reason for summing them together is that sometimes there are very few to 0 palindromes in the first couple intervals so it will be best if we consider the values of k all as one together. We will get

$$\sum_{k=0}^{2} \frac{\lambda^k e^{-\lambda}}{k!} = Number\ of\ intervals\ containing\ 0, 1\ and\ 2\ palindrones$$

For the number of intervals containing 3, 4,…,8 palindromes we will evaluate the Poisson probability individually say,

$$P(K = 8) = \frac{\lambda^8 e^{-\lambda}}{8!} = Number\ of\ intervals\ containing\ 8\ palindrones$$

Lasty for intervals containing 9+ palindromes we will evaluate, the following,

$$\sum_{k=9}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = Number\ of\ intervals\ containing\ 9+\ palindrones$$

After all of this is done we will now have all of our observed and expect numbers for uniformly random and Poisson distributions. The last statistical test we need to examine is the goodness-of-fit for both distributions. We will be using the Chi-square test, this is denoted by,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, O_i = Observed\ result, E_i = Expected\ result$$

A small χ^2 will tell us that the observed and expect results are similar or close together and a large χ^2 tells us that there is a large deviation from the observed and expected result. We want to see which distribution, uniformly random or Poisson has a smaller χ^2. The final statistical

measurement that will be made is that of the p-value. We can obtain the p-value by integrating the χ^2 distribution with 6 degrees of freedom. For the uniformity cases we will be using $(Number\ of\ Intervals - 2)$ degrees of freedom, similar to what we had done for the Poisson distribution. The χ^2 distribution is given by the following,

$$f(x) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{\frac{-x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

$$k = Degrees\ of\ freedom$$

Where,

$$\Gamma(\mathrm{z}) = \int_0^{\infty} x^{z-1}e^{-x}dx$$

Sometimes denoted as,

$$\Gamma(\mathrm{n}) = (n-1)!$$

This means that,

$$\int_0^{\chi^2} f(x)dx = p - value$$

Results

Location of Palindromes	
Mean	116,200
Standard Deviation	64,288
Skewness	-0.0235795
Kurtosis	1.86522
Minimum	117
Maximum	227,316
Median	117,826
Q1	63,549
Q3	170,988

Figure 1. Descriptive Statistics for the locations of palindromes

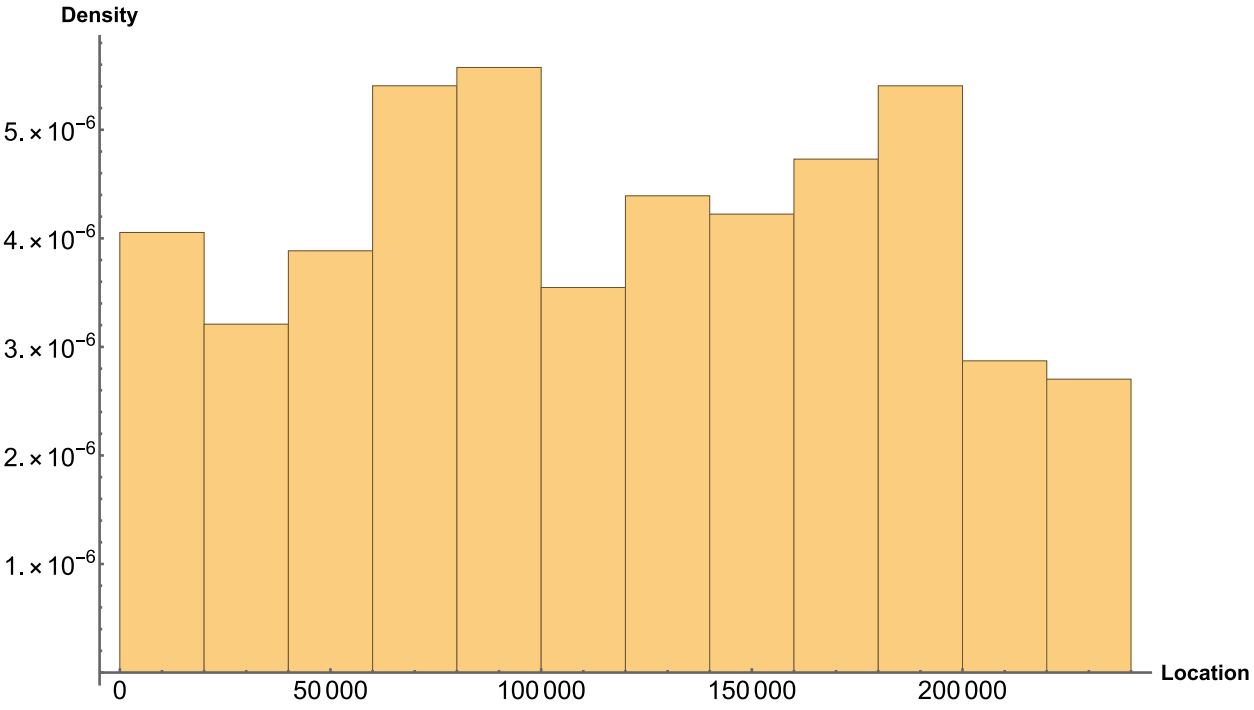

Figure 2. Histogram of Palindrome Locations

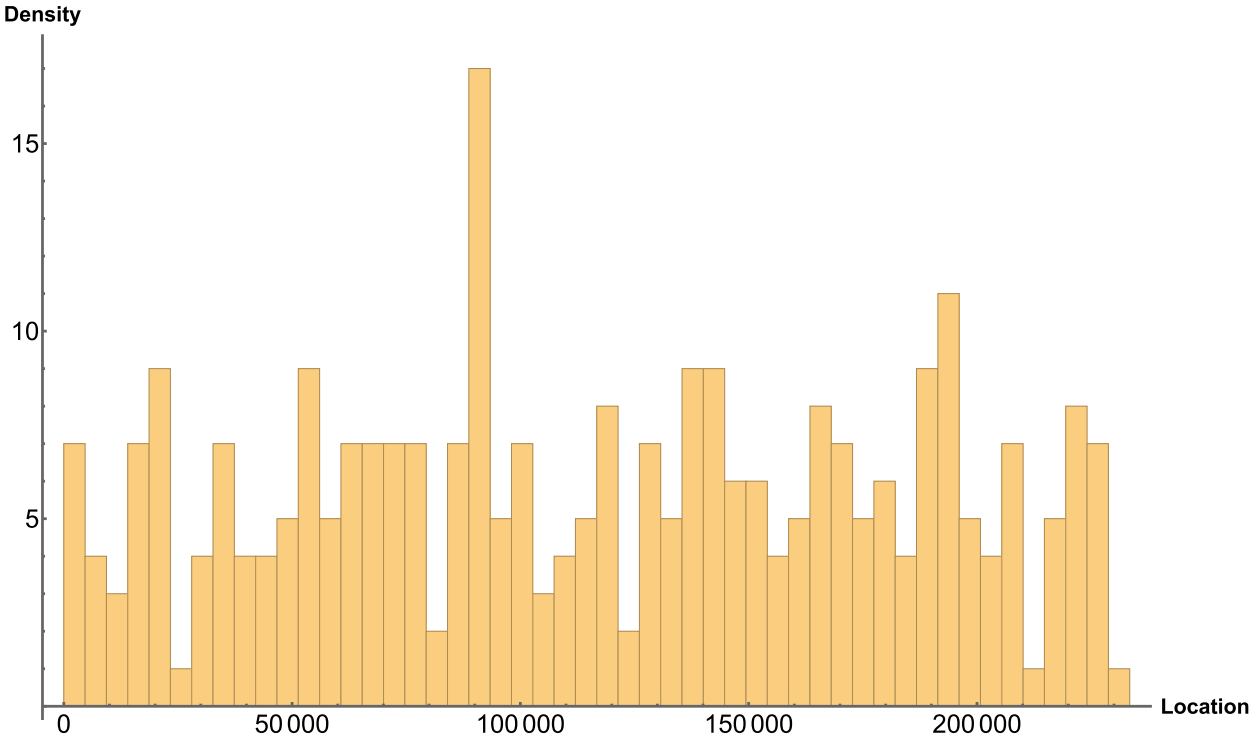

Figure 3. Histogram of Palindrome Locations Using 50 Bins

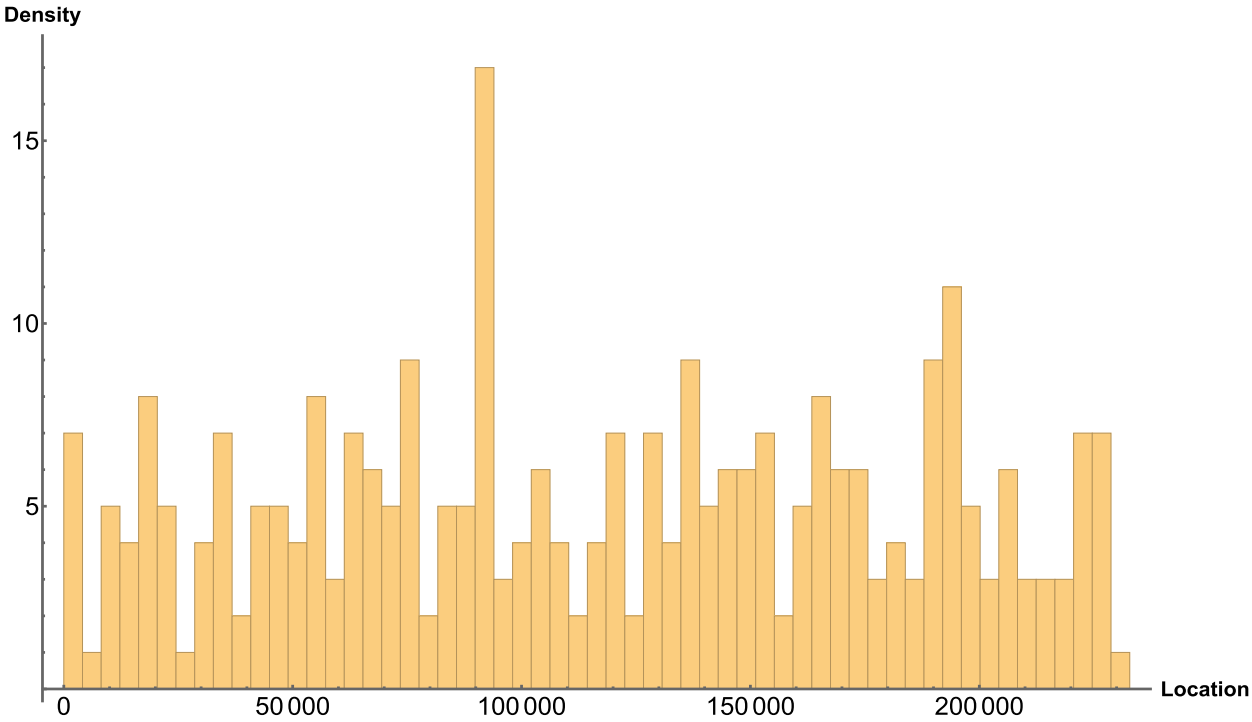

Figure 4. Histogram of Palindrome Locations Using 57 Bins

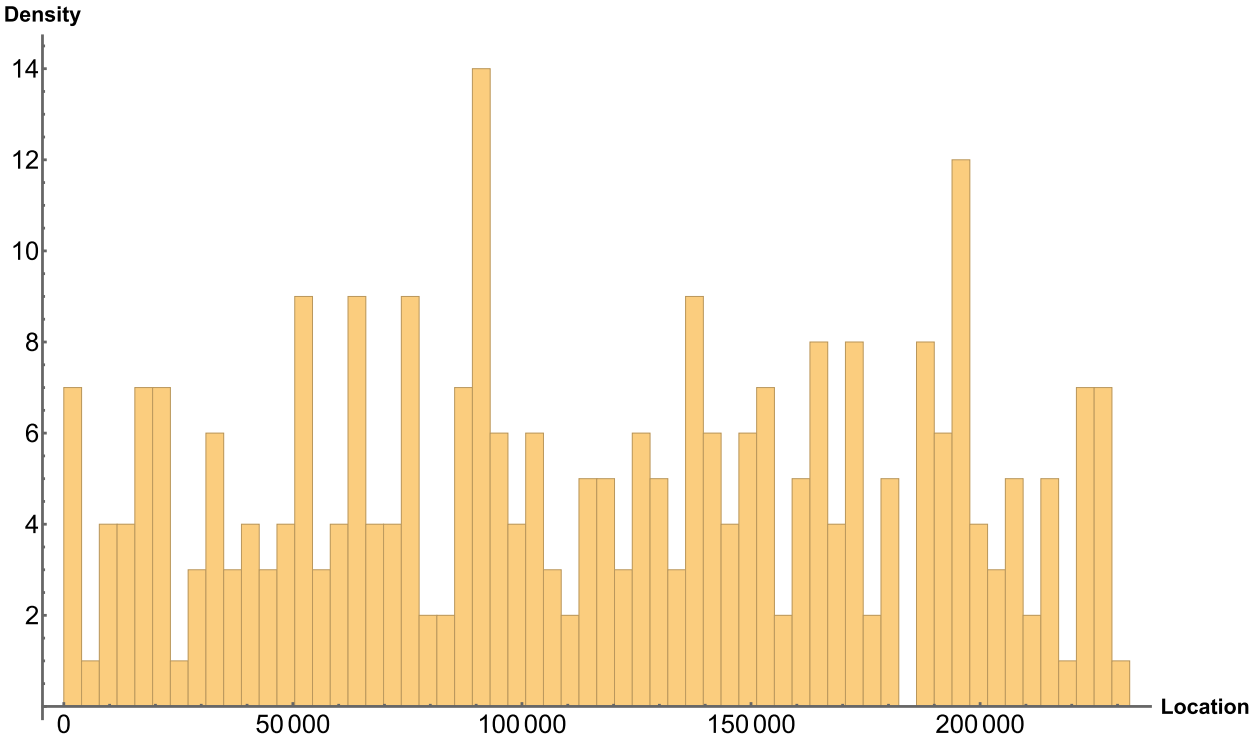

Figure 5. Histogram of Palindrome Locations Using 60 Bins

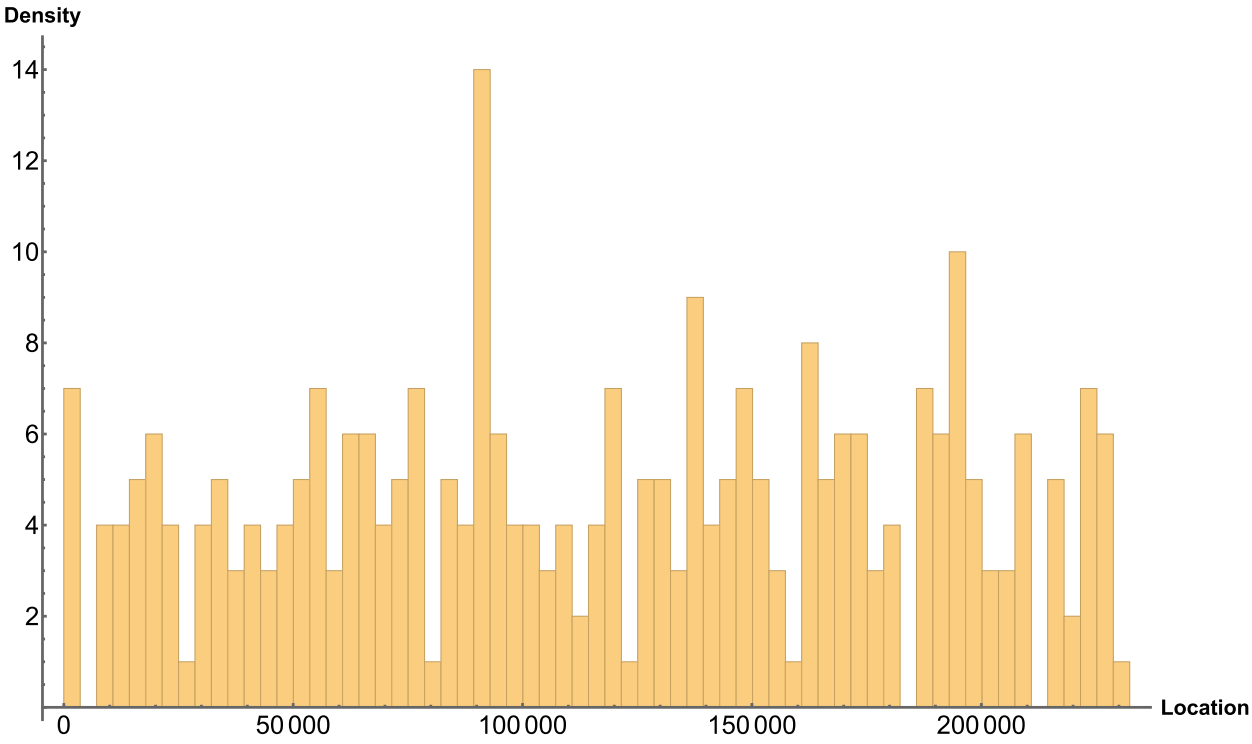

Figure 6. Histogram of Palindrome Locations Using 65 Bins

Due to the large number of observed and expected values for uniformity the entire data set can be found in the reference section of this report.

Palindrome	Number of Intervals	
Count	Observed	Expected
0-2	3	3.37711
3	5	4.73477
4	6	6.96012
5	8	8.1851
6	10	8.0214
7	10	6.73797
8	2	4.95241
9+	6	7.03113
Total	50	50

Figure 7. Poisson Distribution of Observed and Expected Palindromes for 50 Intervals

Segment	1	2	…	57	Total:
Observed	7	1	…	6	294
Expected	5.16	5.16	…	5.16	~294

Figure 8. Uniform Distribution of Observed and Expected Palidromes for 50 Intervals

Palindrome	Number of Intervals	
Count	Observed	Expected
0-2	7	6.38218
3	8	7.5006
4	10	9.67182
5	9	9.97725
6	8	8.57693
7	5	6.31984
8	4	4.07464
9+	6	4.49674
Total	57	57

Figure 9. Poisson Distribution of Observed and Expected Palindromes for 57 Intervals

Segment	1	2	…	50	Total:
Observed	7	1	…	6	294
Expected	5.88	5.88	…	5.88	~294

Figure 10. Uniform Distribution of Observed and Expected Palidromes for 57 Intervals

Palindrome	Number of Intervals	
Count	Observed	Expected
0-2	10	3.37711
3	6	4.73477
4	14	6.96012
5	8	8.1851
6	8	8.0214
7	8	6.73797
8	3	4.95241
9+	3	7.03113
Total	60	60

Figure 11. Poisson Distribution of Observed and Expected Palindromes for 60 Intervals

Segment	1	2	…	60	Total:
Observed	7	1	…	6	294
Expected	4.9	4.9	…	4.9	294

Figure 12. Uniform Distribution of Observed and Expected Palindromes for 60 Intervals

Palindrome	Number of Intervals	
Count	Observed	Expected
0-2	14	11.1149
3	9	10.8822
4	11	12.3053
5	8	11.1315
6	12	8.39146
7	7	5.42217
8	0	3.06561
9+	4	2.68684
Total	65	65

Figure 13. Poisson Distribution of Observed and Expected Palindromes for 65 Intervals

Segment	1	2	…	65	Total:
Observed	7	0	…	6	294
Expected	4.5	4.5	…	4.5	~294

Figure 14. Uniform Distribution of Observed and Expected Palindromes for 65 Intervals

Parameters	Intervals	50	57	60	65
	Degrees of Freedom	48	55	58	63
Chi-Squared Statistic	Uniform	51.5782	74.3371	84.7755	100.5
Area	Uniform	0.664289	0.957766	0.987512	0.998128

Figure 15. Uniform Distribution, χ^2 Statistic and Area (P-Value) Calculations

Parameters	Intervals	50	57	60	65
	Degrees of Freedom	6	6	6	6
Chi-Squared Statistic	Poisson	4.1722	1.01826	3.92615	7.81213
Area	Poisson	0.346614	0.0150894	0.31333	0.747807

Figure 16. Poisson Distribution, χ^2 Statistic and Area (P-Value) Calculations

Discussion & Conclusions

As we can see can in Figure 15 and Figure 16 the Poisson model gives a smaller χ^2 statistic and area (p-value as discussed before) than the Uniform model gives us. At a first-order approximation it would seem that the Poisson model is a better fit for our distribution than the uniform model. Our goal of this report is determining how these palindromes are distributed, did they follow a Poisson model, more uniformly distributed or some other distribution. A secondary goal of this report was to indicate an area or point where DNA replication begins. It is best to examine the histograms in Figures 3-6 to assess if Poisson is truly the better fit, as what our test show. We split the histogram of locations in different bins corresponding to the number of intervals used. When we did this an apparent outlier or two becomes revealed. From our statistical test we believe that the locations followed a Poisson distribution rather than

uniformity, but these different histograms may tell us something otherwise. It would seem that if we were to remove those outliers the histogram would follow a more uniform distribution. This doesn't mean that our tests are wrong, in fact they are right but one of the secondary goals of this report was the assess the location of DNA replication. It may be hypothesized that those outliers are the locations of DNA replication. Those specific points also seem to be very non-uniform which disturbed all statistical tests for uniformity and is why we got the χ^2 values that we did. From these ideas we may think that all other locations are uniformly randomly, and the specific location of replication is not random. This leads us to two possibly conclusions, we will consider and assess both of them.

Our first possible conclusion is that the palindromes do follow a Poisson distribution and it is inconclusive on where DNA replication beings. Our evidence for this relies on the χ^2 statistic and the corresponding p-values. When we compare the χ^2 statistic between the Poisson and uniform cases, Poisson outperformed in every case. Poisson χ^2 statistics are always much smaller than that of the uniforms χ^2 statistics. This leads to the conclusion that the Poisson distribution has a better goodness-of-fit than the uniform case would give us. When using the χ^2 distribution to find the models p-value, Poisson generally has more accepted values, to accept the null hypothesis than uniform does. In this case our null hypothesis would be that the location of palindromes follows a Poisson distribution. The large p-values for uniform tells us that our locations are more likely not uniform.

The next possible conclusion is that the palindromes are uniformly distributed, and the location of DNA replication can be approximately located. Here we will be looking at Figures 3-6 much more. We can see in the histograms that there is a very apparent outlier. When we examine the rest of the data, excluding the outlier, the majority of the data seems to be relatively uniform in a sense. Our hypothesis, discussed before, was that if we were to remove the outlier, our statistical tests would lead to a more conclusive answer toward the locations following uniformity. An issue that arises here is finding out which locations we want to remove and how to assess that those were the correct location we want to remove. From a second-order approximation we will say that the locations follow a uniform distribution, and the location of DNA replication occurs at the outlying data presented in the histograms of differing sized bins.

References

Nolan, D. A., & Speed, T. (2001). Stat labs: Mathematical statistics through applications. New York, NY: Springer.

```
{7, 3, 4, 7, 7, 3, 4, 5, 4, 6, 4, 9, 3, 9, 7, 5, 9, 2, 7,
 5, 16, 5, 6, 4, 2, 6, 6, 6, 5, 6, 8, 7, 7, 6, 3, 8, 7, 5,
 6, 6, 3, 9, 11, 5, 4, 6, 2, 5, 7, 7}
```

Observed Data for 50 Equal Non-overlapping Segments

```
{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 }
```

Observed Data for 57 Equal Non-overlapping Segments

```
{7, 1, 4, 4, 5, 7, 3, 2, 4, 5, 4, 4, 4, 4, 7, 4,
 7, 6, 3, 9, 4, 1, 6, 5, 16, 3, 5, 4, 3, 2, 5, 7,
 1, 7, 3, 6, 7, 8, 4, 5, 5, 2, 5, 8, 6, 6, 3, 4,
 1, 7, 6, 12, 4, 2, 6, 2, 4, 1, 8, 6}
```

Observed Data for 60 Equal Non-overlapping Segments

```
{7, 0, 4, 3, 6, 6, 4, 1, 4, 5, 2, 4, 4, 4, 3, 7, 2, 7, 7, 3, 4, 9, 2, 1, 6, 5, 16, 2,
 4, 5, 3, 2, 3, 4, 6, 2, 7, 3, 5, 6, 9, 1, 6, 5, 4, 2, 6, 6, 6, 5, 3, 4, 0, 7, 6, 10,
 5, 3, 3, 6, 0, 5, 1, 7, 6}
```

Observed Data for 65 Equal Non-overlapping Segments