# MATH 189 Case Study 3

## Patterns in CMV DNA

Yuxin Zou, A13996888
Yifei Li, A92082060
Bo Hu, A13805750
Bo Zhang, A13616488
Xuanyu Wu, A13569778

## Literature Review

DNA, or deoxyribonucleic acid, is the hereditary material which contains the genetic instructions that presents in all forms of known life. Inside DNA are four types of nitrogen-containing chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). There are about 3 billion bases in a human's dna, and altogether represents the uniqueness of each individual, though more than 99 percent of those bases are the same in all people (Newman.) However, since the discovery by Francis Crick and James Watson in 1951, this double-helix structure set up a milestone in modern molecular biology, and had never fail to amaze us with new finding of it and brought us numerous benefits as we apply it to many fields. In agriculture, some DNA-based technologies were used to promote efficiency of crop breeding program, increased the quality and outputs of the product, and protected the eco-environment and so on (Fang). In animal husbandry, the application of genetic biotechnologies such as transfer of cloned genes and recombinant DNA contributed in animal health management, genetic improvement of livestock species and more(RM). However, the biggest credit of DNA goes to the field of pharmaceuticals and medicine for its significance in helping diagnose genetic diseases. In this paper, we will see a research on DNA palindromes that aim to diagnose the way in which CMV virus replicates.

## Introduction and Data Description

The human cytomegalovirus(CMV) is a common virus that is spread through body fluids such as saliva, tears, blood, urine, breast milk, semen, and vaginal fluids. Its easiness to spread and ability to weaken the immune system of the infected patients make it a potentially life-threatening disease. In order to better understand the virus, scientists conducted a research on DNA to study the way in which the virus replicates. The data reflects the search of a special place on the virus' DNA that contains instructions for its reproduction: origin of replication. To find the origin of replication, DNA is cut into segments and each segment is tested to determine its ability to replicate. They found an efficient way of simplifying the process to find the origin of replication-to identify unusually dense clusters of palindromes.

One way to identify the clusters is to divide the DNA chain into multiple intervals and count the number of palindromes in each interval. The data file(hcmv) contains the location of unusual clusters of complementary palindromes within the 296 palindromes that were found at least 10 letters long(they chose to ignore the palindromes that are shorter than 10 letters). The column of the data represents the location that the palindromes are found.

# Investigation:

## Random Scatter

In order to test out if there is significant clusters of palindromes in certain locations, we first visualize the data by grouping them into bins and count the number of palindromes appeared in the range included in each bin. For comparison, we generated another simulation which assumes there is no big clusters and all palindromes placed uniformly among all locations. We do this under the assumption that the appearance of palindromes would follow a poisson process, in which the location of the palindromes would follow discrete uniform distribution. Therefore, we draw 296 numbers from a discrete uniform distribution of length 229354.

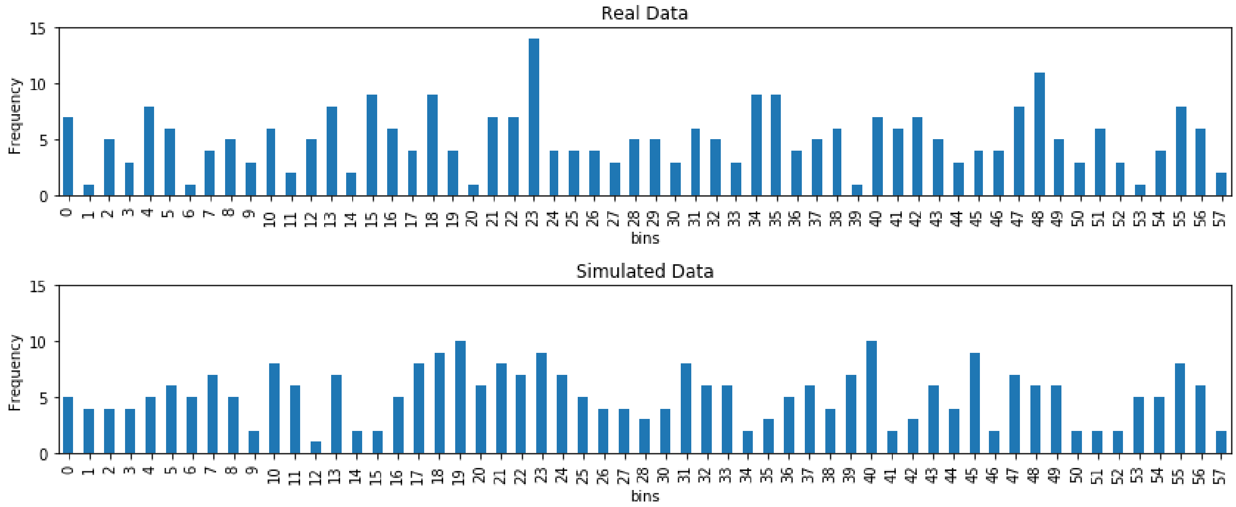


*Figure 1.1: Palindrome location distribution on bins*

By looking at the data that are given, we could see that there is the peak at 24th bin (bin #23), which should alert us that the palindromes' locations might not be uniformly distributed, and it’s likely that there exists places where more palindromes clustered. As a result, further formal analysis will be conducted in the part ‘Counts’.

## Location and spacing

Furthermore, we plot the histograms for the pair distance and the triplets distance among those spacing.

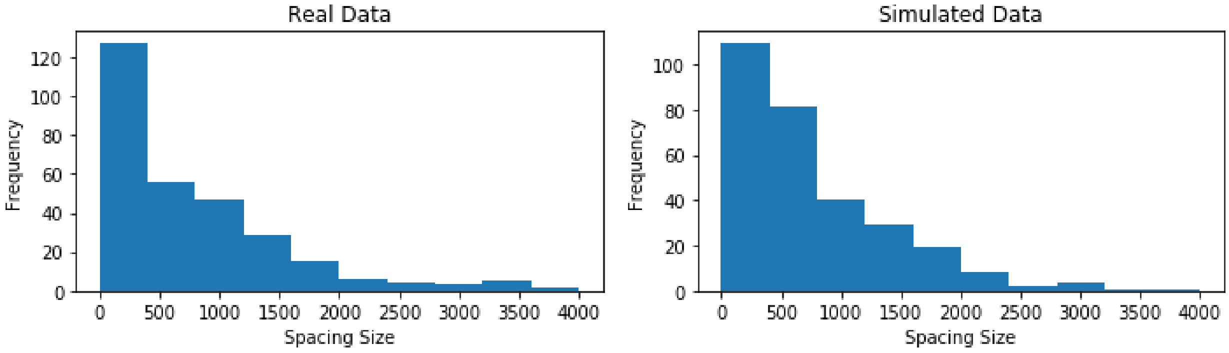


*Figure 1.2: Distribution of Spacing for Pairs*

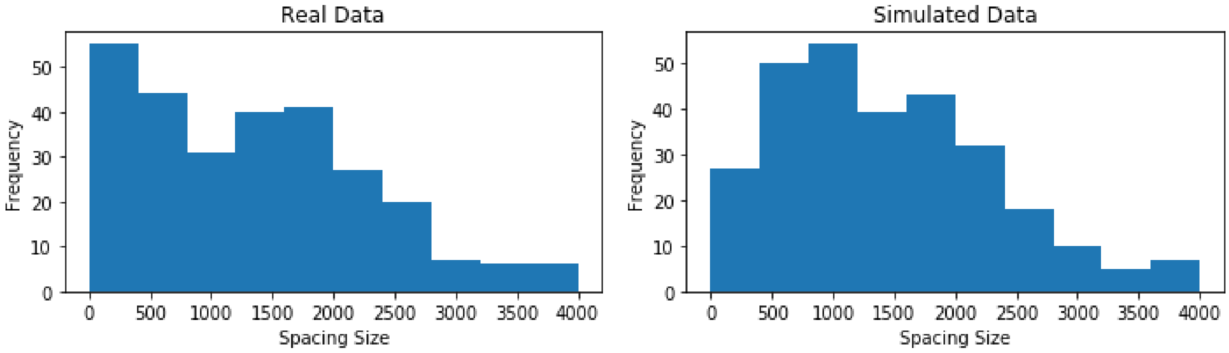


*Figure 1.3: Distribution of Spacing for Triplets*

We could see an exponential distribution clearly, for both the data and the simulation, which means there are lots of palindromes that are very close to each other and forms cluster, while a few of them are further away and isolated. To find out more,  because we spotted a peak at 24th bin from Figure 1.1, we plot the 24th bin's pair distance in a box plot and randomly selected another bin's pair distance boxplot for comparison.

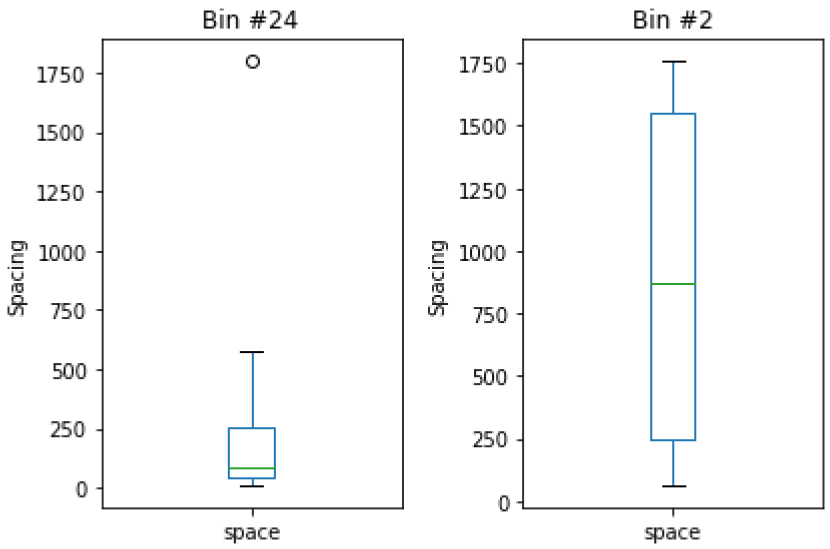


*Figure 1.3: Boxplot for spacing inside bin #24 and bin #2*

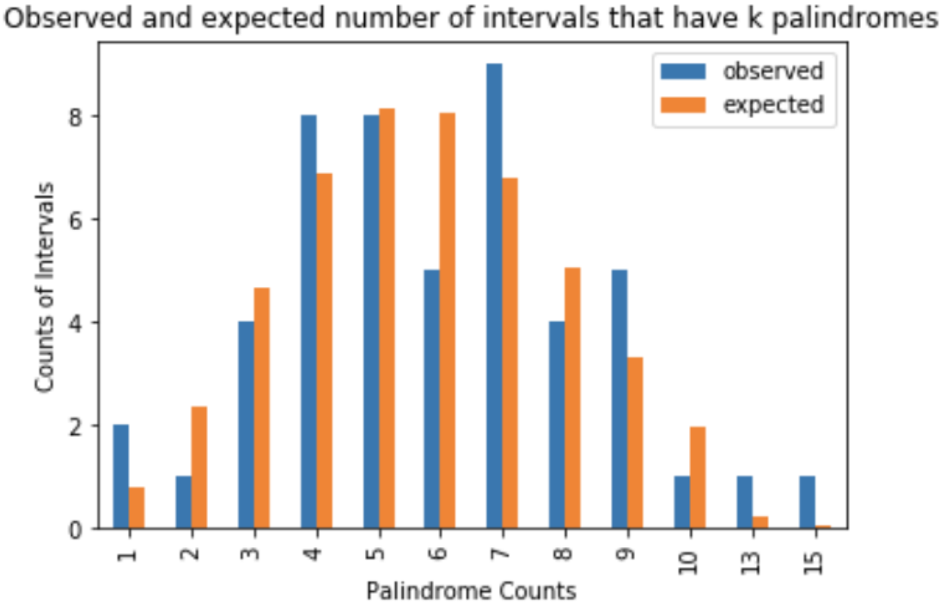
We could see that at the 24th bin, the boxplot is squeezed at the bottom, while the other bins stretched significantly compared to the 24th one. This means that the distances between two palindromes in this location are much shorter compared to other locations, which makes sense because there are more palindromes in this location. If the length of all the bins are fixed, then distances between palindromes in this bin would be much shorter given more palindromes in this location.

## Counts

To examine the counts of palindromes in various regions of the DNA, we split the DNA into 50 non-overlapping regions of the same length so the length of each region is 229354 / 50 = 4587. In each interval, the number of palindromes is recorded. If there is no cluster, the occurrence of palindromes on the DNA will follow a Poisson process. Then, number of palindromes in each interval will follow a Poisson distribution with parameter $\boldsymbol{\lambda}$. Also, palindromes will be scattered randomly and uniformly across the DNA, so counts in these intervals will follow discrete uniform distribution. Therefore, it is helpful that we perform chi-square goodness-of-fit test in both scenarios.

First, we test whether number of intervals that have k palindromes in each interval follows Poisson distribution. The null hypothesis is it follows a Poisson distribution, and the alternative hypothesis is it does not follow a Poisson distribution. In order to estimate $\boldsymbol{\lambda}$, we will use maximum likelihood estimator (MLE), which is the average count in each interval. In this case, since there are 296 palindromes and 50 intervals, $\hat{\lambda} = \frac{296}{50} = 5.92$ . Therefore, the probability of observing k palindromes in an interval is $\frac{\hat{\lambda}^k e^{-\hat{\lambda}}}{k!}$ , and it follows that the expected number intervals that contain k palindromes is $50 * \frac{\hat{\lambda}^k e^{-\hat{\lambda}}}{k!}$ . To get the observed number of intervals that contain k palindromes, we count the appearance of k palindromes in intervals, as shown in Figure 2.1. As we can see, expected and observed counts of intervals are close for each palindrome counts.

*Figure 2.1: Observed and Expected Number of Intervals that Have k Palindromes*



After merging expected numbers that are less than 5 into one cell to make sure that all expected numbers are all greater than 5, Table 2.1 summarizes the expected and observed counts of k-palindrome intervals.

*Table 2.1: Observed and Expected Number of Intervals that Have k Palindromes*

| Palindrome Count | 0-3 | 4 | 5 | 6 | 7 | 8 | 9+ |
|---|---|---|---|---|---|---|---|
| Observed | 7 | 8 | 8 | 5 | 9 | 4 | 8 |
| Expected | 7.92 | 6.87 | 8.14 | 8.03 | 6.79 | 5.02 | 7.23 |

Now, each expected number is greater than 5, so we can compute the test statistic $\sum_i \frac{(Expected_i - Observed_i)^2}{Expected_i}$ =2.45. Using degree of freedom 7-1-1=5, where the first -1 is from the theory, and the second -1 is from the fact that we use estimated $\hat{\lambda}$ instead of true $\lambda$. The p value we get is 0.78, which is larger than our significance level 0.05. Therefore, we fail to reject the null hypothesis, concluding that number of palindromes in each interval seems to follow Poisson distribution.

Next, we want to test whether the number of palindromes in each interval across the DNA follows discrete uniform distribution. The null hypothesis is it follows discrete uniform distribution, and the alternative hypothesis is it does not follow discrete uniform distribution.

Since the DNA is split into 50 equal subintervals, we would expect each interval to contain 5.92 palindromes. Figure 2.2 illustrates the distribution of observed and expected palindrome counts for 50 intervals. We can see that observed and expected counts are close except for 21st and 43rd intervals. So we proceed doing a hypothesis testing.

*Figure 2.2: Observed and Expected Palindrome Counts for 50 Intervals*

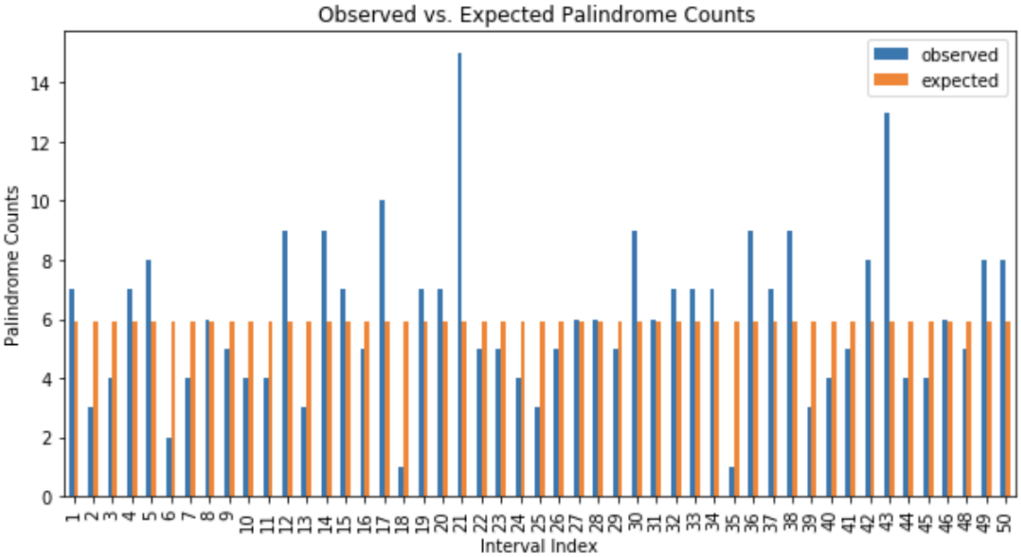


*Table 2.2: Palindrome Counts for 50 Intervals*

| Palindrome Counts |
|---|
| 7 3 4 7 8 2 4 6 5 4 4 9 3 9 7 5 10 1 7 7 15 5 5 4 3 5 6 6 5 9 6 7 7 7 1 9 7 9 3 4 5 8 13 4 4 6 5 8 8 |

The observed palindrome counts in each interval is shown in Table 2.2. Since each expected number is 5.92, which is greater than 5, we can compute our test statistic. The test statistic is $\sum_{i} \frac{(Expected_i - Observed_i)^2}{Expected_i}$ =60.58. Because we did not estimate any parameter, the degree of freedom is 50-1 = 49. The p value is therefore 0.11, which is greater than our significance level of 0.05. Therefore, we fail to reject the null hypothesis. The distribution of number of palindromes in each interval across the DNA is not significantly different from uniform distribution.

Since we fail to reject the null hypothesis in both tests, it appears that Poisson is a reasonable initial model.

## Biggest Cluster

In order to investigate the location of abnormal clusters of palindromes, hypothesis test of a Poisson distribution of locations of palindromes is conducted:

$H_0$: Counts of palindromes follow a Poisson distribution.

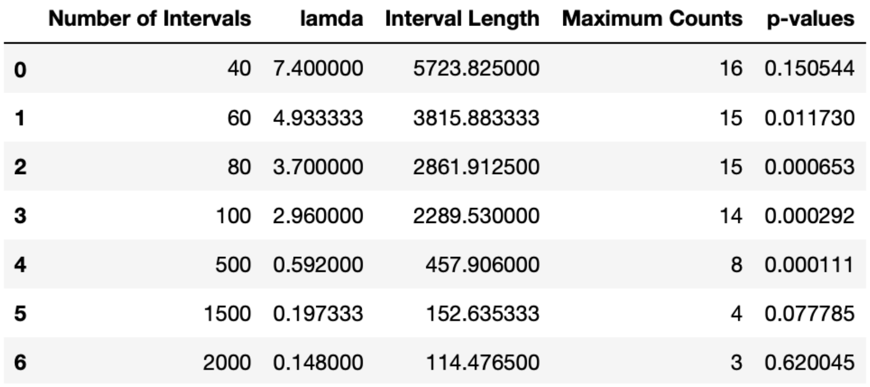
$H_1$: Counts of palindromes are not following a Poisson distribution.

Under the null hypothesis, the number of hits in a set of non-overlapping intervals of the same length are independent observations following a Poisson distribution, which implies that the number of hits of palindromes within each interval are following an identical and independent Poisson Distribution. That is, the greatest number of hits in each of this set of intervals behaves as the maximum of independent Poisson random variables. Based on the i.i.d property of Poisson distributions of number of hits within each interval, the chance that the greatest number of hits is at least k is: (m is the number of intervals; $\lambda$ is the parameter of Poisson distribution.)

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

This chance can be used as the p-value for the test statistic k since an unusual small chance indicates the abnormality of a cluster that is larger than the expectation from the Poisson process with statistical significance. Table 3.1 shown below has seven different number of intervals, which are 40, 60, 80, 100, 500, 1500, 2000, and each has their corresponded $\lambda(lambda)$, interval length, maximum counts over all intervals and the p-value calculated by the formula above. By cutting the entire DNA length into equal-sized intervals, the counts of hits of palindromes can be inferred by determining if a given palindrome falls within this interval or not based on their positions. Since the counts of palindromes follow a Poisson distribution under $H_0$, values of $\lambda(lambda)$ are estimated by Maximum Likelihood Estimator (MLE) of Poisson distribution which is the average number of palindromes within each interval ( Total # of palindromes / Total # of intervals). Interval Length is calculated by dividing total # of intervals from the entire DNA length. Maximum Counts are simply counting the maximum # of hits of palindromes over all intervals.

Table 4.1: P-values of Each Interval Length



|   | Number of Intervals | lamda | Interval Length | Maximum Counts | p-values |
|---|---|---|---|---|---|
| 0 | 40 | 7.400000 | 5723.825000 | 16 | 0.150544 |
| 1 | 60 | 4.933333 | 3815.883333 | 15 | 0.011730 |
| 2 | 80 | 3.700000 | 2861.912500 | 15 | 0.000653 |
| 3 | 100 | 2.960000 | 2289.530000 | 14 | 0.000292 |
| 4 | 500 | 0.592000 | 457.906000 | 8 | 0.000111 |
| 5 | 1500 | 0.197333 | 152.635333 | 4 | 0.077785 |
| 6 | 2000 | 0.148000 | 114.476500 | 3 | 0.620045 |

By setting the significance level $\alpha = 0.05$, we can see that the p-values first decreases below $\alpha$ then increases above $\alpha$ while the interval length narrowing down. The p-values above $\alpha$ indicates that the chance of observing this set of data under $H_0$ is not unusual and we cannot reject $H_0$, and vice versa. The pattern of changes of p-values complies with our common sense. When the interval length is big, unusual clusters of replication sites may merge together with other ones, thus unable to be detected. When we first narrow down the interval length by increasing the total # of intervals, the unusual cluster first become well detected with 60 intervals and a p-value of 0.011730. As we further narrow down the interval by increasing the total # of intervals to 1500, we can see that the p-values once again increases above $\alpha$, leading us unable to reject $H_0$. This phenomenon is reasonable because when the interval length becomes too small, the replicated sites will be split into many consecutive intervals, each containing a very small number of hits close to each other, thus making the distribution similar to Poisson. Usually, a smaller interval length is better in that it tells more information by giving a more accurate position of the unusual cluster. However, if there are too many intervals, the test statistics will be decreased. Thus, we choose the # of intervals of 80 to make a compromise. With 80 intervals, the size of interval is narrowed down to 2861 and we find the location of unusual cluster to be (91581, 94444). For the row of 80 intervals, the p-values is 0.000653, which indicates that the maximum counts of palindromes over these 80 intervals is statistically significantly unusual, meaning that (91581, 94444) has a high potential of being a replication site.

## Advanced Analysis

From investigations of locations and spacing, we notice that the distribution of consecutive spacings looks kind of similar to the exponential distribution, but we cannot be sure if they are statistically close by eyeballing the graph. Therefore, it is essential that we do a chi-square goodness-of-fit test to find out if there is a significant difference between the two distributions. The null hypothesis is that the distribution of spacings is exponential distribution, while the alternative hypothesis is that it is not exponential.

First, we need to estimate the parameter in the exponential distribution. Since the parameter $\lambda$ is the same as the $\lambda$ in the Poisson process, we have $\hat{\lambda} = \frac{296}{229354} = 0.013$ by MLE. Taking each bin size to be 400, we count the number of spacings in each bin as our observed values. To get expected numbers, we multiply the total number of spacings by the probability that a spacing falls into each bin using exponential cumulative distribution function (cdf.) Merging bins to make sure that each bin has expected number of spacings greater than 5, Figure 3 and Table 3 summarized expected and observed values in each bin. We can see that expected and observed counts look close to each other.

*Figure 3: Number of Spacings in Each Interval*

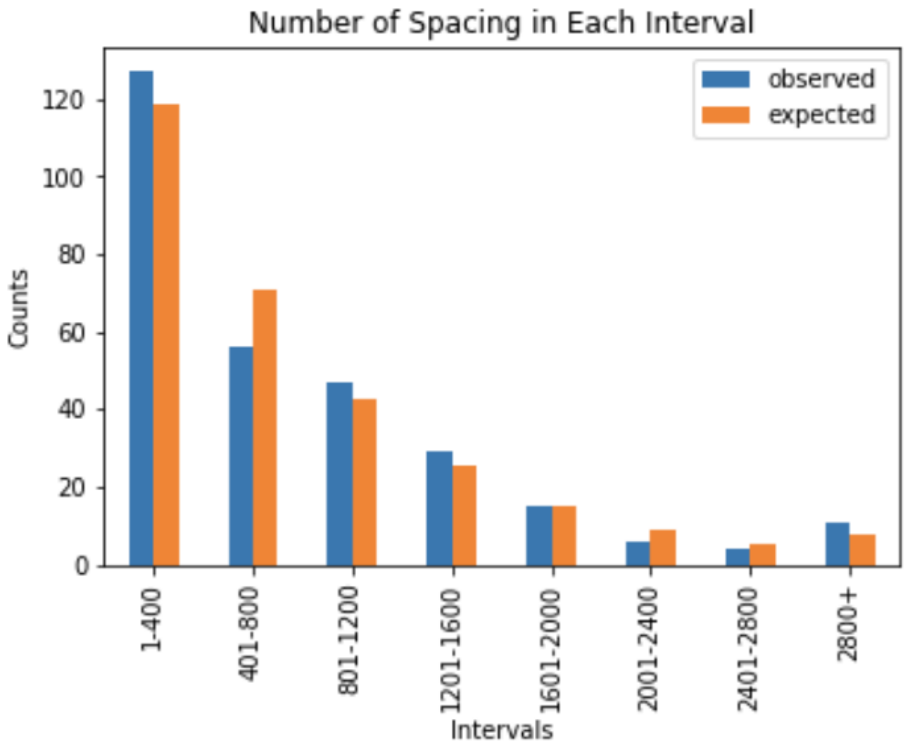


*Table 3: Observed and expected number of spacings that fall into each interval*

| Spacing values | 1-400 | 401-800 | 801-1200 | 1201-1600 | 1601-2000 | 2001-2400 | 2401-2800 | 2801+ |
|---|---|---|---|---|---|---|---|---|
| Observed counts | 127 | 56 | 47 | 29 | 15 | 6 | 4 | 11 |
| Expected counts | 118.95 | 70.99 | 42.36 | 25.28 | 15.07 | 9.00 | 5.37 | 7.95 |

As each expected count is greater than 5, we now compute the test statistic, which is $\sum_i \frac{(Expected_i - Observed_i)^2}{Expected_i}$ =8.28. With degree of freedom 8-1-1=6, this leads to a p value of 0.22. Therefore, we fail to reject the null hypothesis, and there is no significant difference between the two distributions. Thus, the result reconfirms our previous conclusion that Poisson is a reasonable initial model.

## Further Discussion and Conclusion

In the plots of Random Scatter, we could clearly spot a peak at the 24th bin, indicating that a departure from a uniform distribution occurs at the 24th bin. Even though such departures occur, the hypothesis test still fails to reject the null hypothesis that the number of palindromes in each interval follows a Poisson process. That is, it is concluded that Poisson is a reasonable initial model. The fit of Poisson process indicates that palindromes are randomly scattered across a strand of DNA, which have little help in finding the replication sites. Thus further analysis is done with respect to the greatest number of hit across all intervals. And we find that when splitting into 80 intervals, location (91581, 94444) has the abnormal largest cluster with 15 cluster with statistic significance. Thus, it is recommended to the biologist to study location (91581, 94444) for replication site.

## Method

Maximum Likelihood Estimator: Suppose we have an independent sample x1, . . . , xn from a Poisson distribution with unknown rate parameter λ. For Poisson distribution, the chance of observing x1,... , xn is $\frac{\lambda^{x_1}e^{-\lambda}}{x_1!} * ... * \frac{\lambda^{x_n}e^{-\lambda}}{x_n!} = L(\lambda)$ . Since the function is monotonically increasing, the log likelihood function is maximized at the same value as L. Solving the first-order equation $\frac{\partial}{\partial\lambda}L(\lambda) = \frac{\partial}{\partial\lambda}[\sum_i x_i log(\lambda) - n\lambda - \sum_i x_i log(x_i!)] = \sum_i x_i/\lambda - n = 0$ , we obtain $\hat{\lambda} = \bar{x}$. This is the maximum likelihood estimator of $\lambda$ for a Poisson distribution.

## Theory

**P-values:**

In hypothesis testing, a p-value is a number between 0 and 1 that indicates the following: A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, that is, reject the null hypothesis.A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, that is, fail to reject the null hypothesis. *P*-values very close to the cutoff (0.05) are considered to be marginal (could go either way).

**Chi-square Goodness of Fit Test:**

Chi-squared goodness of fit test Chi-squared goodness of fit test is used to determine whether the sample data matches the hypothesized distribution. In our advanced analysis, we expected the distribution of consecutive spacing to be exponential distributed from our observation on the graph. To prove our judgment, we use Chi-square goodness of fit test to find out if there is a significant difference between the two distributions.

**Poisson Process:**

Poisson process is used to count the occurrences of some targeted events that appear to happen in a certain rate by in a random timing, in a given interval of time. It can also be used to model random points in given space and time. In poisson process, events occur continuously and independently at a constant average rate.

**Exponential Distribution:**

An exponential distribution is the probability distribution that describes the time between events in a Poisson point process, which is used to count the occurrences of some targeted events that appear to happen in a certain rate by a random timing, in a given interval of time.

**Discrete Uniform Distribution:**

A discrete uniform distribution is a symmetric probability distribution that contains number of values that have equal chances of being evaluated. It is a known, finite number of outcomes equally likely to happen.

**Poisson Distribution:**

Poisson Distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.

**Degree of Freedom:**

The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called number of degrees of freedom. In other words, the number of degrees of freedom can be defined as the minimum number of independent coordinates that can specify the position of the system completely.

**Null Hypothesis**

Null hypothesis is a status that there is no relationship between two targeted subjects.

The statement being tested in a test of statistical significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually, the null hypothesis is a statement of 'no effect' or 'no difference'. It is often symbolized as $H_0$.

**Alternative hypothesis**

Comparing to null hypothesis, alternative hypothesis is to test a statement against the null hypothesis. It is the hypothesis used in hypothesis testing that is contrary to the null hypothesis. The null hypothesis and the alternative hypothesis are types of conjectures used in statistical tests, which are formal methods of reaching conclusions or making decisions on the basis of data.

## Work cited:

Newman, Tim. “What is DNA and how does it work?” MedicalNewsToday, 11 January 2018, https://www.medicalnewstoday.com/articles/319818.php

Fang, Jinggui  A., et al. “Applications of DNA Technologies in Agriculture.” *Curr Genomics*, PMC, August 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4955036/

RM, Petters. “Recombinant DNA, gene transfer and the future of animal agriculture.” *J Anim Sci*, Pubmed, June 1986, https://www.ncbi.nlm.nih.gov/pubmed/3525489